SHARP NORM CONDITIONS FOR LOGIT FIXED POINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Entropy regularization is often used in multi-agent learning to reduce instability caused by agents influencing one another. A natural intuition is that learning should be easier to stabilize when a change in one agent's behavior has only a limited effect on the others. In this paper, we show that the size of a payoff change alone can be misleading: even a large change may leave an agent's choices unchanged. What matters is whether the change alters the relative payoffs of different actions. Building on this observation, we derive a new sufficient condition for stability of logit responses that accounts only for changes that affect choice, and specifies how much regularization is enough. Under this condition, responses converge to the same fixed point from every initial state. A two-action example establishes the sharpness of the condition's universal constant: immediately beyond the bound, the system can have several distinct long-term outcomes. Experiments show that bounds based on the full payoff matrix can count changes that leave choices unchanged, requiring unnecessarily strong regularization or failing to guarantee convergence in systems that do converge. The new condition avoids this problem. We believe these findings help explain stability in multi-agent learning and guide the choice of regularization strength.

1 INTRODUCTION

Learning to make decisions is a central problem in artificial intelligence. In reinforcement learning, a policy determines how an agent acts and what rewards it receives. When several agents learn together, each agent's decisions also change the rewards available to the others (Mertikopoulos & Sandholm, 2016; Sun et al., 2024). This creates a feedback loop: agents adjust their policies to improve their rewards, but those adjustments change the problem they are trying to solve. Understanding whether the agents eventually stop changing their policies is a basic question in learning with interacting agents.

Entropy regularization offers a way to control this feedback. It favors randomized policies and replaces an abrupt choice of the best action with a smooth response to rewards (Neu et al., 2017; Geist et al., 2019). With Shannon entropy, this response is the familiar softmax, also called the logit response. By making each agent less sensitive to changes in rewards, regularization can help stabilize learning.

Experiments with smooth Q-learning nevertheless show different outcomes. In network zero-sum games, Leonardos et al. (2021) observe convergence to a unique logit equilibrium when all learning agents explore; their theory permits arbitrarily small positive exploration rates. In network Shapley games, Hussain et al. (2023) observe persistent cycles at low positive exploration rates. Increasing exploration eventually restores convergence in these examples, but lowers the agents' total payoff. Thus exploration can resolve instability in one setting while leaving it intact in another, and increasing it can change the quality of the outcome.

These observations make the amount of regularization a substantive part of the learning problem. Sufficient regularization must account for the interaction between agents, while an overly conservative choice can make their decisions unnecessarily insensitive to rewards. Existing analyses relate sufficient regularization to payoff sensitivity (Gao & Pavel, 2017; Cianfanelli & Como, 2025; Sun et al., 2024). We ask: *How much regularization is sufficient to make the agents converge to a unique logit equilibrium?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

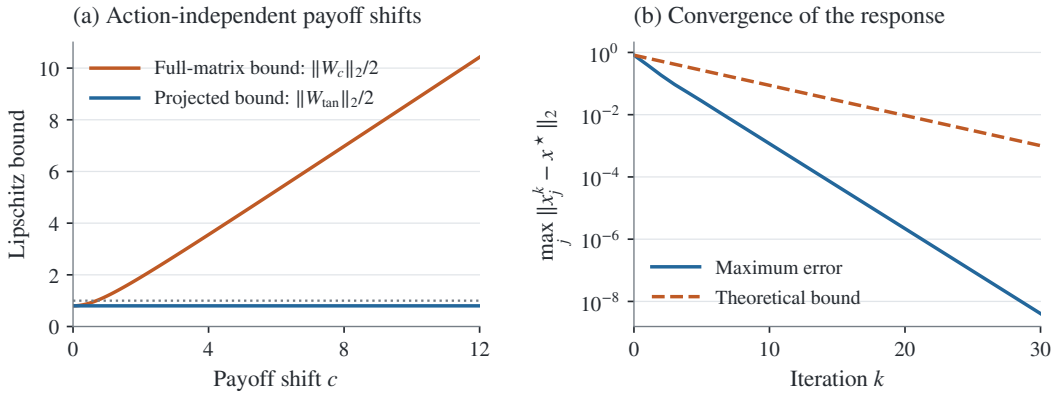


Figure 1: The projected condition remains valid when action-independent payoff shifts make the full-matrix bound fail. (a) For the shifted payoff matrix W_c in Section 4.1, the full-matrix bound $\|W_c\|_2/2$ grows with shift size c, while the projected bound remains 0.8; the dotted line marks one, below which the bound guarantees contraction. (b) Maximum error over 64 initializations and the envelope $0.8^k \max \|x^0 - x^\star\|_2$. Here $x^\star$ is the fixed point. The response map is identical for all c.

We study this question in finite-dimensional logit systems, where a distribution determines payoffs and the logit response determines the next distribution. The key observation is that changes in payoffs need not imply changes in decisions. For example, a payoff shift that is identical across all actions of a given agent leaves its response unchanged, even if the shift depends strongly on other agents' policies. We refer to these as action-independent payoff shifts. A bound that counts this change can therefore become arbitrarily large for an unchanged response map. We remove payoff components that are identical across actions and measure only how feasible changes in probabilities alter payoff differences.

For affine payoffs, we prove uniqueness and global convergence whenever $\beta\|W_{\tan}\|_2 < 2$, where β is inverse temperature and $W_{\tan}$ is the payoff matrix projected onto the simplex tangent space. This projection removes changes shared by all actions. The condition makes the response map a contraction: each update reduces the distance between any two initial distributions. Combining the sharp softmax Lipschitz constant $1/2$ (Nair, 2026) with $\|W_{\tan}\|_2$ gives rates for repeated responses and continuous adjustment. The constant 2 cannot be increased in a universal sufficient condition based only on this norm. In a two-action example, one fixed point becomes three as inverse temperature passes 2: even smooth individual responses can sustain different collective outcomes.

The contraction factor also bounds convergence rates and the error caused by sampling. We obtain linear convergence of independent natural policy gradient (NPG) with exact evaluation in finite games, under an explicit condition on the entropy-regularization coefficient and step size. The required regularization strength can be bounded using either the derivative of the payoff map or the numbers of actions (Section 3.4). When responses are estimated from samples, we quantify how the contraction factor affects the error: diminishing steps give convergence, while constant steps leave an error bound that decreases with batch size.

The analysis extends to multiple populations, nonlinear payoffs, and other regularizers, with a complementary potential argument for symmetric interactions. Synthetic experiments examine the convergence conditions and sampling bounds. The appendices provide complete proofs and experimental protocols. Figure 1 illustrates why removing action-independent payoff shifts matters.

2 Related Work

Regularization and learning dynamics. Entropy regularization is studied in control and reinforcement learning (Kappen, 2005; Todorov, 2006; 2009; Ziebart et al., 2008; Neu et al., 2017; Geist et al., 2019). Convergence analyses cover regularized policy gradients and NPG (Cen et al., 2022; Ding et al., 2025; Sun et al., 2024). For smooth Q-learning, Leonardos et al. (2021) establish convergence in weighted zero-sum polymatrix games, while Hussain et al. (2023) study the effect

of exploration on convergence and payoff performance in network games. Dynamical systems and stochastic approximation provide further tools for analyzing regularized responses and stochastic learning (Hofbauer & Sandholm, 2002; Fudenberg & Levine, 1998; Benaïm, 1996; 1999; Benaïm et al., 2005; 2006; Mertikopoulos & Sandholm, 2016).

Quantal response equilibrium. A quantal response equilibrium is a strategy profile in which each player uses a probabilistic response to the expected payoffs under that same profile (McKelvey & Palfrey, 1995; 1998; Goeree et al., 2008; 2016). Related population and social-interaction models examine the emergence of multiple equilibria (Blume, 1993; Brock & Durlauf, 2001; Alós-Ferrer & Netzer, 2010; Sandholm, 2010; Haile et al., 2008; Melo, 2022). The logit response is a particular probabilistic choice rule. Although it assigns a unique distribution to each payoff vector, a game can have several logit equilibria. This distinction also motivates convergence analyses of population logit dynamics (Cianfanelli & Como, 2025) and regularized mean-field iterations (Cui & Koeppl, 2021).

Lipschitz continuity of softmax. Gao & Pavel (2017) developed convex-analytic and monotone-operator properties of softmax and used them in learning dynamics, often using a Lipschitz upper bound of 1 for softmax. Nair (2026) established the sharp constant $1/2$ across ℓ_p norms. We apply this Lipschitz bound to the composition of the payoff map and softmax. Projecting the payoff matrix removes directions that do not affect the response; the binary example then proves optimality of the universal constant in the resulting norm condition.

Uniqueness at high temperature. The competition between entropy and interaction also arises in Gibbs measures and variational inference (Dobruschin, 1968; Georgii, 2011; Georgii et al., 2006; Mezard & Montanari, 2009; Jordan et al., 1999; Wainwright & Jordan, 2008). Dobrushin criteria bound absolute influences between blocks (Dobrushin & Shlosman, 1985; Weitz, 2005; Hayes, 2006; Martinelli, 1999); related analyses study uniqueness in Bethe and message-passing formulations (Yedidia et al., 2005; Heskes, 2004; Mooij & Kappen, 2007). Absolute-influence bounds can exploit sparse or triangular structure. Our Euclidean condition retains the signs of interactions and can exploit cancellation. We compare it with a Dobrushin-type block contraction condition based on blockwise Lipschitz bounds. A Hadamard construction separates these two criteria; neither uniformly dominates the other.

3 THEORY

Theorems 1, 2, and 3 give conditions for global convergence, uniqueness under symmetry, and convergence of independent NPG. We develop their extensions in the same order. Operator norms are Euclidean unless specified otherwise; the spectral norm is the largest singular value.

3.1 A SHARP CONDITION FOR GLOBAL CONVERGENCE

We first study a single distribution over actions. Systems with several players require one distribution per player and are introduced in Section 3.2.

Let x describe the probabilities of $n \geq 2$ actions, so x belongs to the simplex $\Delta = \{x \in \mathbb{R}^n_{\geq 0} : \mathbf{1}^\top x = 1\}$. The payoff vector is $Wx + b$, where $W \in \mathbb{R}^{n\times n}$ is the payoff matrix and $b \in \mathbb{R}^n$ is a fixed payoff vector. For inverse temperature $\beta > 0$ and softmax $\sigma_i(z) = e^{z_i} / \sum_j e^{z_j}$, a fixed point satisfies

$$x = F_\beta(x) := \sigma\big(\beta(Wx + b)\big), \qquad x \in \Delta, \tag{1}$$

The distribution can also describe population shares or variational marginals. We study repeated responses, $x^{k+1} = F_\beta(x^k)$ (Picard iteration), and continuous adjustment, $\dot{x} = F_\beta(x) - x$. A fixed point of F_β is a rest point of this ODE. In population games it is called a logit equilibrium (Cianfanelli & Como, 2025); Section 3.4 gives the corresponding definition for finite games. Continuity and compactness ensure that a fixed point exists. We seek conditions for uniqueness and global convergence.

Projection onto the simplex tangent space. Probability can move between actions, but its total stays one, so feasible changes sum to zero. Also, $\sigma(z + c\mathbf{1}) = \sigma(z)$: a payoff change shared by every action has no effect on choice. These observations give

$$\mathcal{T} = \mathbf{1}^{\perp}, \qquad \Pi = I - \mathbf{1}\mathbf{1}^{\top}/n, \tag{2}$$

$$W_{\tan} = \Pi W \Pi|_{\mathcal{T}}. \tag{3}$$

Here $\mathcal{T}$ is the simplex tangent space and Π is its orthogonal projection, which subtracts the mean. We call $W_{\tan}$ the projected payoff matrix. The projections restrict changes to feasible directions and remove action-independent payoff shifts. Thus $\|W_{\tan}\|_2$ measures the largest change in centered payoffs per unit change in the distribution. Replacing W by $W + \mathbf{1}a^{\top}$ leaves $W_{\tan}$ and the response unchanged. Adding $c\mathbf{1}^{\top}$ leaves $W_{\tan}$ unchanged and only changes the bias on Δ.

The remaining step is to bound the response to relative payoffs. For $p = \sigma(z)$, the derivative is the covariance matrix

$$D\sigma(z) = \Sigma(p) := \mathrm{Diag}(p) - pp^{\top}, \qquad \Sigma(p)\mathbf{1} = 0, \qquad \|\Sigma(p)\|_2 \le \tfrac{1}{2}. \tag{4}$$

The constant is attained at the uniform binary distribution; for more than two actions it is approached by concentrating mass equally on two coordinates. The derivative identity is classical, and the sharp bound across ℓ_p norms is due to Nair (2026). Its Euclidean form has a simple interpretation. For a unit vector v,

$$v^{\top}\Sigma(p)v = \mathrm{Var}_{i\sim p}(v_i) \le \frac{(\max_i v_i - \min_i v_i)^2}{4} \le \frac{1}{2}. \tag{5}$$

Here $(v_i - v_j)^2 \le 2(v_i^2 + v_j^2)$. The largest response occurs when probability is split equally between two actions and the perturbation favors one over the other. This explains the binary extremizer; Appendix A.1 gives the complete proof.

Global convergence. A change in the distribution first changes relative payoffs and then changes the response. The product $q = \beta\|W_{\tan}\|_2/2$ is an upper bound on the Lipschitz constant of F_β: $\|F_\beta(x) - F_\beta(y)\|_2 \le q\|x - y\|_2$. When $q < 1$, it is a contraction factor, so every update brings any two distributions closer together.

Theorem 1. *Let $q = \beta\|W_{\tan}\|_2/2 < 1$. Then F_β is a global q-contraction on Δ and has a unique fixed point $x^\star$. For every $x^0 \in \Delta$, Picard iteration obeys*

$$\|x^k - x^\star\|_2 \le q^k\|x^0 - x^\star\|_2, \qquad \|x^k - x^\star\|_2 \le \frac{\|x^{k+1} - x^k\|_2}{1 - q}. \tag{6}$$

The ODE $\dot{x} = F_\beta(x) - x$ has a unique global solution in Δ for every initial condition there, enters its relative interior for $t > 0$, and satisfies $\|x(t) - x^\star\|_2 \le e^{-(1-q)t}\|x(0) - x^\star\|_2$. The constant 2 in $\beta\|W_{\tan}\|_2 < 2$ is optimal among universal sufficient conditions based only on $\|W_{\tan}\|_2$.

Proof of convergence. For a feasible direction $h \in \mathcal{T}$, $h = \Pi h$ and $\Sigma(p)\Pi = \Sigma(p)$. Hence

$$DF_\beta(x)h = \beta\Sigma(F_\beta(x))\Pi W \Pi h, \qquad \|DF_\beta(x)|_{\mathcal{T}}\|_2 \le q. \tag{7}$$

Integrating this derivative along the segment between $x, y \in \Delta$ gives $\|F_\beta(x) - F_\beta(y)\|_2 \le q\|x - y\|_2$. Banach's theorem gives the unique fixed point and the Picard rate; summing the contraction tail gives the residual bound in (6).

For continuous adjustment, Appendix A.1 establishes global existence and positivity. The convergence rate follows from the same contraction bound: with $e = x - x^\star$,

$$\frac{1}{2}\frac{d}{dt}\|e\|_2^2 = \langle e, F_\beta(x) - F_\beta(x^\star)\rangle - \|e\|_2^2 \le -(1 - q)\|e\|_2^2. \tag{8}$$

Gronwall's inequality gives the stated rate. □

No symmetry, monotonicity, or sign assumptions on W are needed. Equation (6) also gives a stopping criterion: the change between successive iterates bounds the remaining error.

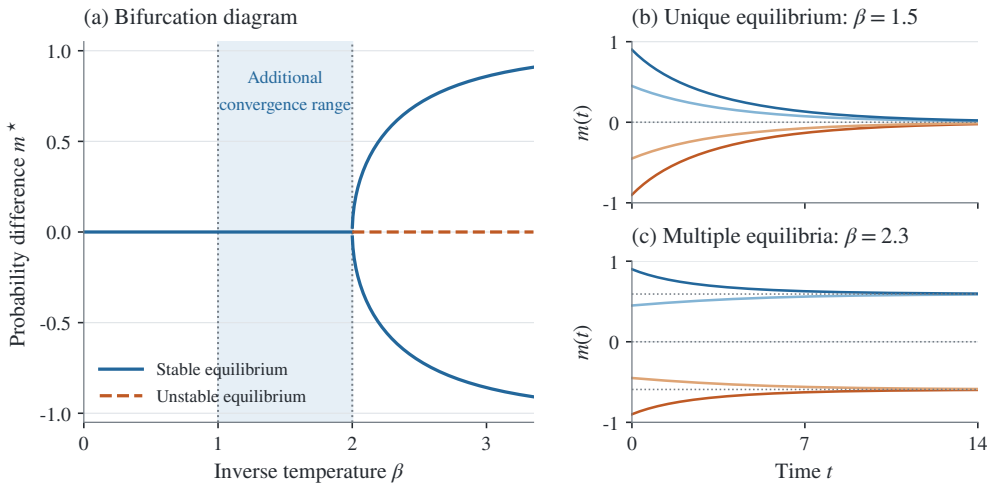


Figure 2: **Bifurcation diagram for the two-action logit dynamics.** Here $m = x_1 - x_2$. (a) Fixed points of $m = \tanh(\beta m/2)$: solid branches are stable and the dashed branch is unstable. Shading marks $1 \le \beta < 2$, where our condition guarantees convergence but the unit Lipschitz bound does not. (b–c) Continuous adjustment converges to the same state below 2 and depends on initialization above 2.

A two-action example. Can the universal constant be increased? Consider two actions that become more attractive as their own probability increases. Entropy opposes this reinforcing effect; the example identifies where uniqueness is lost. For $W = \left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$ and $b = 0$, $\|W_{\tan}\|_2 = \lambda_{\max}(W_{\tan}) = 1$. In the coordinate $m = x_1 - x_2$, the fixed-point equation is

$$m = \tanh(\beta m/2). \tag{9}$$

There is exactly one fixed point for $\beta \le 2$ and exactly three for $\beta > 2$. Consequently, no constant $C > 2$ can make $\beta\|W_{\tan}\|_2 < C$ a universal sufficient condition for contraction or uniqueness. This proves sharpness in Theorem 1; Appendix A.1 gives the root-counting argument. At $\beta = 2$, this family still has a unique fixed point, but no strict contraction. The norm does not describe every individual system: $W = -cI$ has a unique potential maximizer for every $c, \beta > 0$, even when $\beta\|W_{\tan}\|_2$ is arbitrarily large. Uniqueness in that case does not imply Picard contraction.

The binary model describes a population or a stateless learner with two actions. Its ODE is

$$\dot{m} = \tanh(\beta m/2) - m. \tag{10}$$

For $\beta < 2$, the right side has sign opposite to every nonzero m, and $|m(t)| \le e^{-(1-\beta/2)t}|m(0)|$. At $\beta = 2$, the rest point remains unique and attracting, but the linearized decay rate vanishes. Above 2, the origin is unstable and the two nonzero rest points are attracting. The value $\beta = 2$ is a bifurcation point for this family. Our condition covers the entire interval below it, including $1 \le \beta < 2$, which the unit Lipschitz estimate misses. Figure 2 shows this change in the fixed points and adjustment trajectories; Appendix C.4 gives the numerical procedure.

3.2 Interacting populations and nonlinear payoffs

The same argument applies when several agents choose their policies simultaneously: probability is conserved within each policy, and only payoff differences within that agent affect its response. For m populations or players, let

$$\mathcal{X} = \Delta_1 \times \cdots \times \Delta_m, \qquad \mathcal{T} = \mathcal{T}_1 \oplus \cdots \oplus \mathcal{T}_m, \tag{11}$$

where $\Delta_a \subset \mathbb{R}^{n_a}$ and $\mathcal{T}_a = \{v^a : \mathbf{1}^\top v^a = 0\}$. Let $\Pi = \mathrm{blkdiag}(\Pi_1, \ldots, \Pi_m)$ be the block tangent projection. With block matrices $W^{ab} \in \mathbb{R}^{n_a \times n_b}$ and block biases b^a, define

$$F^a(x) = \sigma_a\left(\beta_a\left(b^a + \sum_{c=1}^{m} W^{ac}x^c\right)\right), \qquad a = 1, \ldots, m, \tag{12}$$

where each block has its own inverse temperature $\beta_a > 0$. We write

$$B_\beta = \mathrm{blkdiag}(\beta_1 I_{n_1}, \dots, \beta_m I_{n_m}). \tag{13}$$

A block represents population shares or a player's mixed strategy. In polymatrix games, payoffs are sums of pairwise interactions and $W^{aa} = 0$. General finite games may have nonlinear payoff maps and are covered by the extension below. Population models may have nonzero diagonal blocks. Blocks with only one action can be eliminated.

Applying the argument of Theorem 1 within each block gives the condition

$$q_{\mathrm{prod}} = \tfrac{1}{2}\|B_\beta \Pi W \Pi|_{\mathcal{T}}\|_2 < 1. \tag{14}$$

Under this condition, F has a unique fixed point, Picard iteration contracts with factor q_{prod}, and block logit adjustment converges with rate $e^{-(1-q_{\mathrm{prod}})t}$ from every point of $\mathcal{X}$. For a probability profile $p = (p^1, \dots, p^m) \in \mathcal{X}$, define

$$\Sigma_{\mathrm{blk}}(p) = \mathrm{blkdiag}\big(\Sigma(p^1), \dots, \Sigma(p^m)\big).$$

The Jacobian factors through this block covariance matrix:

$$DF(x) = \Sigma_{\mathrm{blk}}(F(x)) B_\beta W, \qquad DF(x)h = \Sigma_{\mathrm{blk}}(F(x)) B_\beta \Pi W \Pi h \quad (h \in \mathcal{T}). \tag{15}$$

Each covariance block annihilates its all-ones direction, and B_β is scalar on every block. A block-diagonal matrix has norm equal to the maximum of the block norms, so

$$\|DF(x)|_{\mathcal{T}}\|_2 \le \tfrac{1}{2}\|B_\beta \Pi W \Pi|_{\mathcal{T}}\|_2. \tag{16}$$

Integrating along segments proves contraction; Appendix A.2 also gives the ODE argument. The softmax constant remains $1/2$, so dependence on the number of populations enters through the projected matrix norm.

Comparison with a Dobrushin-type condition. Our condition takes the norm of the full projected payoff matrix. A Dobrushin-type block condition instead replaces each interaction block by a nonnegative Lipschitz bound, which can lose cancellations between interactions. Appendix B.3 constructs systems with two actions per population and $\beta_a = 1$ for every population. In this family, our bound remains fixed while the block bound grows with the square root of the number of populations. Triangular interactions can give the reverse comparison, so neither criterion uniformly dominates the other.

Nonlinear payoffs. When the effect of a policy change depends on other policies, the fixed matrix W is replaced by the derivative of a nonlinear payoff map G. A uniform bound on its projected derivative gives a global result. Let G be C^1 on a neighborhood of Δ, and define

$$L_T = \sup_{x \in \Delta} \|\Pi DG(x) \Pi|_{\mathcal{T}}\|_2. \tag{17}$$

Thus L_T is a Lipschitz bound for the centered payoff map $x \mapsto \Pi G(x)$ on Δ. For $F(x) = \sigma(\beta G(x))$, the uniqueness and convergence conclusions of Theorem 1 hold with $q = \beta L_T/2 < 1$. For $F = \sigma_{\mathrm{blk}}(B_\beta G)$ on $\mathcal{X}$, it suffices that

$$q_G = \frac{1}{2} \sup_{x \in \mathcal{X}} \|B_\beta \Pi DG(x) \Pi|_{\mathcal{T}}\|_2 < 1. \tag{18}$$

The uniform derivative bound makes the affine result usable when interaction varies with the distribution. For example, Appendix A.2 proves the computable bound $L_T \le \|\Pi W \Pi|_{\mathcal{T}}\|_2 + \|\Pi C\|_2 \|A \Pi\|_2$ for $G(x) = Wx + b + C\tanh(Ax + d)$. This separates the affine effect from the nonlinear correction. A bound restricted to a smaller forward-invariant region may be sharper than the global one.

Other norms and regularizers. The contraction argument extends to every ℓ_p norm, where softmax retains Lipschitz constant $1/2$, and to uniformly strongly convex regularizers. The chosen norm determines how action-independent payoff shifts are removed; the regularizer's convexity modulus determines response sensitivity. Appendices B.1–B.2 define the corresponding quotient norms and give the conditions and proofs.

Table 1: Sufficient affine conditions. Contraction gives uniqueness and convergence of iteration and continuous adjustment; concavity gives uniqueness through the potential. Let $A_\beta = B_\beta^{1/2}\Pi W \Pi B_\beta^{1/2}|_{\mathcal{T}}$. Projected operators act on the tangent space. The reference bounds use a unit Lipschitz constant for softmax or a unit lower bound on entropy curvature.

Domain	Guarantee	Reference bound	Sharp bound
Single	Contraction	$\beta\|W\|_2 < 1$	$\beta\|W_{\tan}\|_2 < 2$
Product	Contraction	$\|B_\beta W\|_2 < 1$	$\|B_\beta \Pi W \Pi|_{\mathcal{T}}\|_2 < 2$
Single	Concavity*	$\beta\lambda_{\max}(W_{\tan}) < 1$	$\beta\lambda_{\max}(W_{\tan}) < 2$
Product	Concavity*	$\lambda_{\max}(A_\beta) < 1$	$\lambda_{\max}(A_\beta) < 2$

*Requires $W = W^\top$.

3.3 Uniqueness under symmetric interactions

When the payoff matrix is symmetric, the logit fixed points are the interior stationary points of an entropy-regularized potential. Strict concavity then gives uniqueness, including in some systems where the response is not a contraction.

Theorem 2. *Suppose $W = W^\top$. Define, on Δ, with $0 \log 0 = 0$,*

$$\Phi_\beta(x) = H(x) + \frac{\beta}{2} x^\top W x + \beta b^\top x, \quad H(x) = -\sum_i x_i \log x_i. \tag{19}$$

If $\mu = 2 - \beta\lambda_{\max}(W_{\tan}) > 0$, then for $x, y \in \Delta$ and $t \in [0,1]$,

$$\Phi_\beta(tx + (1-t)y) \geq t\Phi_\beta(x) + (1-t)\Phi_\beta(y) + \frac{\mu}{2} t(1-t)\|x - y\|_2^2. \tag{20}$$

Its unique maximizer is interior and equals the unique logit fixed point.

The theorem compares entropy and payoff curvature along feasible directions. On Δ°, for $v \in \mathcal{T}$,

$$v^\top \nabla^2 H(x) v = -\sum_i \frac{v_i^2}{x_i}, \qquad \sum_i \frac{v_i^2}{x_i} \geq 2\|v\|_2^2. \tag{21}$$

The second inequality follows by applying Cauchy–Schwarz to the positive and negative parts of the zero-sum vector v. The constant 2 is attained in the uniform binary case. The payoff term adds at most $\beta\lambda_{\max}(W_{\tan})\|v\|_2^2$ to the Hessian, so integrating along segments gives (20). Entropy keeps the maximizer interior, where the KKT equations give the logit fixed-point equation. Appendix A.3 proves the curvature bound and supplies the boundary and stationarity arguments.

For several populations, the same argument weights each population's entropy by its temperature.

For symmetric W, let

$$\mathcal{V}(x) = \sum_a \beta_a^{-1} H(x^a) + \tfrac{1}{2} x^\top W x + b^\top x, \tag{22}$$

If

$$\mu_{\text{prod}} = 2 - \lambda_{\max}(B_\beta^{1/2}\Pi W \Pi B_\beta^{1/2}|_{\mathcal{T}}) > 0, \tag{23}$$

then $\mathcal{V}$ satisfies (20) on $\mathcal{X}$ with Φ_β replaced by $\mathcal{V}$ and the final term replaced by $\frac{\mu_{\text{prod}}}{2} t(1-t)\|B_\beta^{-1/2}(x-y)\|_2^2$. Its unique maximizer is the unique block-logit fixed point. This is the blockwise form of Theorem 2; its proof is included in Appendix A.3.

Table 1 compares contraction and concavity with unit Lipschitz or unit entropy-curvature estimates. The intermediate condition $\beta\|W\|_2 < 2$ isolates the gain from softmax sensitivity before projection.

3.4 Convergence of learning dynamics

We now bound the error from sampled responses and establish linear convergence of independent natural policy gradient. Appendix A.4 contains the proofs.

Sampled responses. When actions are sampled, convergence depends on the balance between contraction and noise. Averaging draws reduces the noise; a step size limits how much the estimate changes the current distribution. Given the history $\mathcal{F}_k$, draw M conditionally independent categorical samples with probabilities $F(x_k)$, and write their empirical distribution as $\widehat{F}_{k+1}$. Consider

$$x_{k+1} = (1 - \alpha_k)x_k + \alpha_k \widehat{F}_{k+1}, \qquad 0 < \alpha_k \leq 1. \tag{24}$$

If F is a q-contraction with $q < 1$, conditional unbiasedness and variance at most $1/M$ give

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2 \mid \mathcal{F}_k] \leq [1 - (1 - q)\alpha_k]^2 \|x_k - x^\star\|_2^2 + \frac{\alpha_k^2}{M}. \tag{25}$$

Robbins–Monro steps ($\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$) imply almost-sure and mean-square convergence. For $\alpha_k = a/(k + k_0)$, $k_0 \geq a$, and $2a(1 - q) > 1$, the mean-square error is $O(k^{-2a(1-q)}) + O(1/(Mk))$. At constant step size α, it instead satisfies

$$\limsup_k \mathbb{E}\|x_k - x^\star\|_2^2 \leq \frac{\alpha}{M(1 - q)[2 - \alpha(1 - q)]}. \tag{26}$$

Equation (26) separates the roles of contraction, sampling, and step size. A smaller contraction factor q lowers the bound; a larger batch reduces it in proportion to $1/M$. Diminishing steps remove the persistent error. With one sample and no averaging over time, every update remains at a simplex vertex and cannot approach an interior fixed point. Appendix A.4 gives the proof and the variance factor m/M for m blocks.

Independent natural policy gradient. Independent NPG updates all players simultaneously. In the finite-game setting of Sun et al. (2024), player i uses policy π_i and exact expected rewards. Write $G_i(\pi)(a_i) = \mathbb{E}_{a_{-i} \sim \pi_{-i}} r_i(a_i, a_{-i})$ for the reward of action a_i, averaged over the other players, and stack these vectors into $G(\pi)$. Here $\tau > 0$ is the entropy-regularization coefficient, corresponding to temperature $1/\beta$. A logit quantal response equilibrium (QRE) satisfies $\pi_i = \sigma_i(G_i(\pi)/\tau)$ for every player. Equivalently, each π_i maximizes $\langle \pi_i, G_i(\pi) \rangle + \tau H(\pi_i)$ with the other players' policies held fixed. It is therefore a Nash equilibrium of the entropy-regularized game. Independent NPG uses the update

$$\pi_{k+1,i}(a_i) \propto \pi_{k,i}(a_i)^{1-\alpha\tau} \exp\{\alpha G_i(\pi_k)(a_i)\}. \tag{27}$$

To see how the response bound applies, take logarithms of this multiplicative update. For policies with full support, let $s = \Pi \log \pi$, which subtracts the mean log probability within each player. Since $\pi = \sigma_{\mathrm{blk}}(s)$, the update becomes a weighted combination of the current log policy and the relative payoffs:

$$s_{k+1} = (1 - \alpha\tau)s_k + \alpha \Pi G(\sigma_{\mathrm{blk}}(s_k)). \tag{28}$$

Theorem 3. *Let $L_T = \sup_{\pi \in \mathcal{X}} \|\Pi D G(\pi) \Pi|_{\mathcal{T}}\|_2$. If $\tau > L_T/2$ and $0 < \alpha\tau \leq 1$, then the centered update* (28) *contracts on $\mathcal{T}$ with factor*

$$\rho = 1 - \alpha\tau + \alpha L_T/2 < 1. \tag{29}$$

There is a unique QRE $\pi^\star$, and every full-support initialization satisfies $\|s_k - s^\star\|_2 \leq \rho^k \|s_0 - s^\star\|_2$ and $\|\pi_k - \pi^\star\|_2 \leq \rho^k \|s_0 - s^\star\|_2/2$. If $r_i \in [0, 1]$ and $n_i = |\mathcal{A}_i|$, then

$$L_T \leq \frac{1}{2}\sqrt{\sum_{i \neq j} n_i n_j}, \qquad \tau > \frac{1}{4}\sqrt{\sum_{i \neq j} n_i n_j} \tag{30}$$

is an explicit sufficient condition on the entropy-regularization coefficient.

The proof differentiates (28), using the covariance factor on its right. Each off-diagonal block of DG has entries in $[0, 1]$; centering and the variance bound give $\|\Pi_i D_j G_i \Pi_j\|_F \leq \sqrt{n_i n_j}/2$. Summing squared Frobenius norms proves (30); Appendix A.4 gives the complete argument. Polymatrix games have a constant, directly computable projected payoff matrix.

Writing $S = \sum_i n_i$, the sufficient threshold on τ in (30) is below $S/4$, versus $2S$ in Theorem III.3 of Sun et al. (2024). Both concern the same exact update with rewards in $[0, 1]$. We require $0 < \alpha\tau \leq 1$ and bound Euclidean distance between centered log policies; their rate uses QRE gap. A tighter bound on L_T further lowers our sufficient threshold.

Computing a sufficient temperature. For affine or polymatrix payoffs, compute $\|Q^\top W Q\|_2$ using an orthonormal tangent basis Q. Nonlinear payoffs require a uniform derivative bound, such as (50) or (30); derivative measurements along finitely many trajectories alone do not give a global guarantee.

The same argument applies to population logit adjustment, discounted score learning, and exact mean-field iteration once the relevant payoff or value sensitivity is bounded. Appendices B.4–B.5 derive these applications, including the settings of Cianfanelli & Como (2025); Cui & Koeppl (2021).

4 Experiments

We examine three consequences of the theory: invariance to action-independent payoff shifts, convergence beyond a unit Lipschitz bound, and the effect of batch size on sampling error. Each construction isolates one effect for comparison with its analytic bound.

4.1 Invariance to payoff shifts

Can a payoff change make a norm bound worse without changing the learning dynamics? We test this in a three-action system by adding a shift of size c that is identical across actions. The projected matrix norm stays at 1.6, giving contraction factor 0.8, while the full matrix norm grows as $\sqrt{1.6^2 + 3c^2}$. Figure 1 shows that the projected condition continues to certify convergence after the full-matrix bound exceeds one.

For all 64 tested initializations, the shifted and unshifted responses agree to numerical precision. The largest observed error ratio is below 0.8. We use this term for the largest ratio $\|x^{k+1} - x^\star\|_2 / \|x^k - x^\star\|_2$ over all initializations and steps whose denominator exceeds 10^{-10}. Appendix C.1 gives the construction and measurements. Thus the full-matrix bound can lose its guarantee even though the dynamics are unchanged.

4.2 Nonlinear and population dynamics

We test the nonlinear payoff map from Section 3.2 and a generalized rock–scissor–paper population game (Cianfanelli & Como, 2025). In both examples, the unit Lipschitz estimate is inconclusive, while the sharp response bound guarantees contraction. All 64 tested trajectories stay below the theoretical error bounds (Figure 3). Appendix C.2 reports the configurations and observed error ratios.

4.3 Stochastic response updates

For the binary system with $\beta = 1.6$, we compare batches $M \in \{1, 4, 16, 64\}$ using 512 trajectories per setting and three seeds. Under both harmonic and constant steps, mean-square error decreases approximately as $1/M$. Every measured constant-step tail error lies below (26). Figure 4 and Table 2 report the errors and theoretical bounds; Appendix C gives the schedules, measurement windows, and fitted slopes.

5 Conclusion

We derived a sharp norm condition for uniqueness and global convergence of logit responses by measuring how policy changes alter relative payoffs. In the binary example, multiple fixed points appear immediately beyond the sufficient bound. Extensions to learning give convergence rates and sampling-error bounds. Together, these results relate sufficient regularization to how changes in policies alter relative payoffs.

Limitations and future work. Our bounds are uniform over the feasible domain; accounting for covariance variation or invariant regions may yield sharper conditions. Nonsmooth payoffs, infinite-dimensional systems, dependent or biased sampling, delayed or asynchronous updates, and function approximation require further analysis. General regularizers require a positive uniform convexity modulus.

AI USE STATEMENT

AI tools assisted literature searches, proof drafting and revision, experimental design, numerical implementation and code editing. The authors are responsible for the correctness, originality, and presentation of the submitted work.

REPRODUCIBILITY STATEMENT

Appendix A provides the proofs of the main results, and Appendix B develops the further extensions. Appendix C specifies the numerical constructions, initializations, random seeds, step sizes, and error measurements. All experiments use synthetic finite-dimensional systems without external datasets.

REFERENCES

Carlos Alós-Ferrer and Nick Netzer. The logit-response dynamics. *Games and Economic Behavior*, 68(2):413–427, 2010.

Michel Benaïm. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer, 1999.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

Lawrence E Blume. The statistical mechanics of strategic interaction. *Games and economic behavior*, 5(3):387–424, 1993.

William A. Brock and Steven N. Durlauf. Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260, 2001.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022.

Leonardo Cianfanelli and Giacomo Como. On the stability of the logit dynamics in population games. *IEEE Transactions on Automatic Control*, 70(9):5910–5925, 2025. doi: 10.1109/TAC.2025.3553096. URL https://arxiv.org/abs/2307.03507.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1909–1917, 2021. URL https://proceedings.mlr.press/v130/cui21a.html.

Yuhao Ding, Junzi Zhang, Hyunin Lee, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *IEEE Transactions on Automatic Control*, 2025.

P. L. Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

R. L. Dobrushin and S. B. Shlosman. Completely analytical gibbs fields. In *Statistical Physics and Dynamical Systems*, volume 10 of *Progress in Physics*, pp. 371–403. Birkhäuser, 1985.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169, 2019.

Hans-Otto Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter, 2 edition, 2011.

Hans-Otto Georgii, Salvador Miracle-Sole, Jean Ruiz, and Valentin A Zagrebnov. Mean-field theory of the potts gas. *Journal of Physics A: Mathematical and General*, 39(29):9045, 2006.

Jacob K. Goeree, Charles A. Holt, and Thomas R. Palfrey. Quantal response equilibria. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 2008.

Jacob K. Goeree, Charles A. Holt, and Thomas R. Palfrey. *Quantal Response Equilibrium: A Stochastic Theory of Games*. Princeton University Press, 2016.

Philip A. Haile, Ali Hortaçsu, and Grigory Kosenok. On the empirical content of quantal response equilibrium. *American Economic Review*, 98(1):180–200, 2008.

Thomas P. Hayes. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 39–46, 2006.

Tom Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16 (11):2379–2413, 2004.

Josef Hofbauer and William H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Aamal Hussain, Francesco Belardinelli, and Dario Paccagnan. The impact of exploration on convergence and performance of multi-agent Q-learning dynamics. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14178–14202, 2023. URL https://proceedings.mlr.press/v202/hussain23a.html.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Hilbert J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005.

Stefanos Leonardos, Georgios Piliouras, and Kelly Spendlove. Exploration-exploitation in multi-agent competition: Convergence with bounded rationality. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26318–26331, 2021. URL https://arxiv.org/abs/2106.12928.

Fabio Martinelli. Lectures on glauber dynamics for discrete spin models. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1997)*, volume 1717 of *Lecture Notes in Mathematics*, pp. 93–191. Springer, 1999.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1(1):9–41, 1998.

Emerson Melo. On the uniqueness of quantal response equilibria and its application to network games. *Economic Theory*, 74(3):681–725, 2022.

Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Joris M Mooij and Hilbert J Kappen. Sufficient conditions for convergence of the sum–product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, 2007.

Pravin Nair. Softmax is $1/2$-lipschitz: A tight bound across all ℓ_p norms. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=6dowaHsa6D.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. arXiv preprint arXiv:1705.07798, 2017.

William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010.

Youbang Sun, Tao Liu, P. R. Kumar, and Shahin Shahrampour. Linear convergence of independent natural policy gradient in games with entropy regularization. *IEEE Control Systems Letters*, 8: 1217–1222, 2024. URL https://arxiv.org/abs/2405.02769.

Emanuel Todorov. Linearly-solvable markov decision problems. *Advances in neural information processing systems*, 19, 2006.

Emanuel Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Dror Weitz. Combinatorial criteria for uniqueness of gibbs measures. *Random Structures & Algorithms*, 27(4):445–475, 2005.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7): 2282–2312, 2005.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A PROOFS OF THE MAIN RESULTS

The proofs follow the order of Section 3: global convergence and sharpness, extensions to several populations and nonlinear payoffs, uniqueness for symmetric interactions, and learning dynamics. We use the notation introduced with each result in the main text.

A.1 GLOBAL CONVERGENCE AND SHARPNESS

We begin with the softmax sensitivity bound used in Theorem 1. The notation is as follows: $\mathbf{1}$ is the all-ones vector, $\mathcal{T} = \{v : \mathbf{1}^\top v = 0\}$ in the single-simplex case, and $\Pi = I - n^{-1}\mathbf{1}\mathbf{1}^\top$ is the orthogonal projection onto $\mathcal{T}$.

Derivation of (4). For $\sigma(z)_i = e^{z_i}/\sum_j e^{z_j}$ and $p = \sigma(z)$,

$$\frac{\partial \sigma_i}{\partial z_j}(z) = p_i(\mathbf{1}\{i=j\} - p_j), \qquad D\sigma(z) = \mathrm{Diag}(p) - pp^\top = \Sigma(p). \tag{31}$$

With $z = \beta(Wx + b)$, the chain rule gives $DF_\beta(x) = \beta\Sigma(F_\beta(x))W$. Since $\mathbf{1}^\top\Sigma(p) = 0^\top$, we have $\mathbf{1}^\top DF_\beta(x) = 0^\top$, so every derivative output lies in $\mathcal{T}$.

For any unit vector v,

$$v^\top\Sigma(p)v = \sum_i p_i v_i^2 - \Big(\sum_i p_i v_i\Big)^2 = \mathrm{Var}_{i\sim p}(v_i) \le \frac{(\max_i v_i - \min_i v_i)^2}{4} \le \frac{\|v\|_2^2}{2} = \frac{1}{2}. \tag{32}$$

The penultimate inequality follows from $(v_i - v_j)^2 \le 2(v_i^2 + v_j^2) \le 2\|v\|_2^2$. Equality is attained by $p = (1/2, 1/2, 0, \ldots, 0)$ and $v = (1, -1, 0, \ldots, 0)/\sqrt{2}$. □

Global contraction and continuous adjustment.

Proof of the contraction statement in Theorem 1. For $h \in \mathcal{T}$, $h = \Pi h$. Also $\Sigma(p)\Pi = \Sigma(p)$ because $\Sigma(p)\mathbf{1} = 0$. Hence

$$DF_\beta(x)h = \beta\Sigma(F_\beta(x))Wh = \beta\Sigma(F_\beta(x))\Pi W\Pi h. \tag{33}$$

By the bound in (4),

$$\sup_{z\in\Delta} \|DF_\beta(z)|_{\mathcal{T}}\|_2 \le \frac{\beta}{2}\|W_{\tan}\|_2 = q. \tag{34}$$

For $x, y \in \Delta$, set $h = x - y \in \mathcal{T}$. The segment $y + th$ remains in Δ, and the integral formula gives

$$F_\beta(x) - F_\beta(y) = \int_0^1 DF_\beta(y + th)h\, dt. \tag{35}$$

Thus $\|F_\beta(x) - F_\beta(y)\|_2 \le q\|x - y\|_2$. If $q < 1$, Banach's fixed-point theorem on the complete metric space Δ gives existence, uniqueness, and the Picard rate. The a posteriori estimate follows by summing the contraction tail:

$$\|x^k - x^\star\|_2 \le \sum_{j=k}^{\infty}\|x^{j+1} - x^j\|_2 \le \sum_{r=0}^{\infty} q^r\|x^{k+1} - x^k\|_2 = \frac{\|x^{k+1} - x^k\|_2}{1-q}. \tag{36}$$

□

Proof of the ODE statement in Theorem 1. The affine softmax vector field is smooth on $\mathbb{R}^n$. Let $x(t)$ be its maximal local solution with $x(0) \in \Delta$. For every time in its existence interval, variation of constants gives

$$x_i(t) = e^{-t}x_i(0) + \int_0^t e^{-(t-s)}F_{\beta,i}(x(s))\, ds \ge 0. \tag{37}$$

Because every softmax coordinate is positive, the inequality is strict when $t > 0$. Moreover,

$$\mathbf{1}^\top x(t) = e^{-t}\mathbf{1}^\top x(0) + \int_0^t e^{-(t-s)}\mathbf{1}^\top F_\beta(x(s))\, ds = 1.$$

Thus the compact simplex is forward invariant, and the continuation theorem extends the solution globally. Let $y(t) = x(t) - x^\star$. For $y(t) \neq 0$,

$$\frac{d}{dt}\|y(t)\|_2 = \Big\langle \frac{y(t)}{\|y(t)\|_2}, F_\beta(x(t)) - F_\beta(x^\star) - y(t) \Big\rangle \leq q\|y(t)\|_2 - \|y(t)\|_2 = -(1-q)\|y(t)\|_2. \tag{38}$$

Gronwall's inequality gives the claimed exponential bound; the case $y(t) = 0$ is immediate. □

Optimality of the constant.

Proof of the optimality assertion in Theorem 1. Write

$$x(m) = \left(\frac{1+m}{2}, \frac{1-m}{2}\right), \qquad m \in [-1, 1]. \tag{39}$$

For $W = \left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$,

$$Wx(m) = \left(-\frac{1-m}{2}, -\frac{1+m}{2}\right), \qquad \beta[(Wx)_1 - (Wx)_2] = \beta m. \tag{40}$$

Thus

$$F_\beta(x(m))_1 - F_\beta(x(m))_2 = \frac{e^{\beta(Wx)_1} - e^{\beta(Wx)_2}}{e^{\beta(Wx)_1} + e^{\beta(Wx)_2}} = \tanh\left(\frac{\beta m}{2}\right), \tag{41}$$

so fixed points are exactly the roots of

$$m = \tanh\left(\frac{\beta m}{2}\right). \tag{42}$$

Let $a = \beta/2$. The equation is odd. If $a \leq 1$, then for $m > 0$,

$$\tanh(am) < am \leq m, \tag{43}$$

so the only root is $m = 0$. If $a > 1$, then $g(m) = \tanh(am) - m$ satisfies $g(0) = 0$, $g'(0) = a - 1 > 0$, and $g(1) = \tanh(a) - 1 < 0$, so a positive root exists. It is unique because $r(m) = \tanh(am)/m$ is strictly decreasing on $(0, 1]$:

$$r'(m) = \frac{am\,\mathrm{sech}^2(am) - \tanh(am)}{m^2} < 0, \tag{44}$$

since $y\,\mathrm{sech}^2 y < \tanh y$ for every $y > 0$, equivalently $y < \sinh y \cosh y$. Oddness gives exactly two nonzero roots $\pm m_\beta$. Finally, $\mathcal{T} = \mathrm{span}\{(1, -1)\}$ and $W(1, -1)^\top = (1, -1)^\top$, hence $\lambda_{\max}(W_{\tan}) = 1$. □

Dynamics of the binary example. In the coordinate $m = x_1 - x_2$, the ODE is $\dot{m} = \tanh(\beta m/2) - m$. For $\beta \leq 2$, its right side has sign opposite to $m \neq 0$, giving global attraction to zero; for $\beta < 2$, $|m(t)| \leq e^{-(1-\beta/2)t}|m(0)|$. At $\beta = 2$, uniqueness and attraction hold but the linearized rate vanishes. For $\beta > 2$, the origin is unstable and the two nonzero roots are attracting, by the sign change of the scalar right side. These are properties of this binary family, not conclusions for arbitrary interaction matrices.

A.2 Interacting populations and nonlinear payoffs

Interacting populations. Let $\mathcal{T} = \oplus_a \mathcal{T}_a$ and $\Pi = \mathrm{blkdiag}(\Pi_1, \ldots, \Pi_m)$. For a probability profile $p \in \mathcal{X}$, recall

$$\Sigma_{\mathrm{blk}}(p) = \mathrm{blkdiag}\big(\Sigma(p^1), \ldots, \Sigma(p^m)\big). \tag{45}$$

Then $\|\Sigma_{\mathrm{blk}}(p)\|_2 \leq 1/2$ because it is block diagonal and each block is a softmax covariance matrix.

Proof of the block contraction bound. For $h \in \mathcal{T}$, $h = \Pi h$. The block Jacobian satisfies

$$DF(x) = \Sigma_{\mathrm{blk}}(F(x)) B_\beta W. \tag{46}$$

Since each covariance block annihilates the corresponding all-ones direction and B_β is scalar on each block,

$$DF(x)h = \Sigma_{\mathrm{blk}}(F(x))B_\beta \Pi W \Pi h. \tag{47}$$

Therefore

$$\|DF(x)|_{\mathcal{T}}\|_2 \le \frac{1}{2}\|B_\beta \Pi W \Pi\|_{\mathcal{T}\to\mathcal{T}} = q_{\mathrm{prod}}. \tag{48}$$

The same integral argument on the convex product simplex gives

$$\|F(x) - F(y)\|_2 \le q_{\mathrm{prod}}\|x - y\|_2. \tag{49}$$

If $q_{\mathrm{prod}} < 1$, Banach's theorem gives a unique fixed point and convergence of Picard iteration. For continuous time, variation of constants applies within each block, proving positive coordinates for $t > 0$ and unit mass in each block. Compactness of $\mathcal{X}$ gives global existence. The single-simplex distance argument then applies with q replaced by q_{prod}. □

Nonlinear payoffs.

Proof of the nonlinear extension. For $h \in \mathcal{T}$, the chain rule and $\Sigma(p)\Pi = \Sigma(p)$ give

$$DF(x)h = \beta\Sigma(F(x))\Pi DG(x)\Pi h.$$

Its norm is at most $\beta L_T\|h\|_2/2$. Since the segment between two simplex points stays in the simplex, integrating this derivative gives the global Lipschitz estimate. The block derivative is $DF(x)h = \Sigma_{\mathrm{blk}}(F(x))B_\beta\Pi DG(x)\Pi h$; the covariance blocks have norm at most $1/2$, proving (18). Banach's theorem gives the discrete-time statements.

For the ODE, G is C^1 on an open neighborhood of the compact domain, hence the vector field is locally Lipschitz there. Consider the maximal local solution. The response remains a probability vector wherever this solution is defined. Variation of constants, as in (37), proves nonnegativity and unit mass up to the maximal time, block by block. Positivity is strict at every positive time. Thus the solution stays in the compact simplex or product simplex, and the continuation theorem rules out finite maximal time. Finally, with $e = x - x^\star$,

$$\frac{1}{2}\frac{d}{dt}\|e\|_2^2 = \langle e, F(x) - F(x^\star)\rangle - \|e\|_2^2 \le -(1-q)\|e\|_2^2,$$

including at $e = 0$. Gronwall's inequality proves the rate. □

For the nonlinear model $G(x) = Wx + b + C\tanh(Ax + d)$, write $D(x) = \mathrm{Diag}(\mathrm{sech}^2(Ax + d))$. Since $\|D(x)\|_2 \le 1$,

$$\|\Pi DG(x)\Pi|_{\mathcal{T}}\|_2 \le \|\Pi W\Pi|_{\mathcal{T}}\|_2 + \|\Pi C\|_2\|A\Pi\|_2. \tag{50}$$

This bound holds uniformly over the domain. Applying the argument on a smaller region requires forward invariance and either the segment property or a Lipschitz bound on that region.

A.3 Uniqueness under symmetric interactions

Throughout this appendix the matrix under discussion is symmetric, and all eigenvalues are computed on the corresponding tangent space.

A single population. On Δ°,

$$\nabla\Phi_\beta(x)_i = -1 - \log x_i + \beta(Wx + b)_i, \qquad \nabla^2\Phi_\beta(x) = -\mathrm{Diag}(1/x) + \beta W. \tag{51}$$

Interior maximizers and KKT equations. Every global maximizer of Φ_β over Δ lies in Δ°. Indeed, if $x_j = 0$ and $x_i > 0$, then for $x(\varepsilon) = x + \varepsilon e_j - \varepsilon e_i$,

$$\Phi_\beta(x(\varepsilon)) - \Phi_\beta(x) = -\varepsilon\log\varepsilon + O(\varepsilon) > 0 \tag{52}$$

for all sufficiently small $\varepsilon > 0$; the quadratic and linear terms contribute only $O(\varepsilon)$.

For $x \in \Delta^\circ$, the KKT condition for maximizing Φ_β over the affine simplex is that $\nabla\Phi_\beta(x) = \lambda\mathbf{1}$ for some $\lambda \in \mathbb{R}$. By (51), this is equivalent to

$$-1 - \log x_i + \beta(Wx + b)_i = \lambda \quad\Longleftrightarrow\quad x_i = \exp(\beta(Wx + b)_i - 1 - \lambda). \tag{53}$$

Normalizing over i gives $x = \sigma(\beta(Wx + b))$. Conversely, the softmax equation gives the KKT condition. Thus fixed points are exactly interior KKT points of Φ_β.

Entropy curvature on the tangent space. For $x \in \Delta^\circ$ and $v \in \mathcal{T}$,

$$v^\top \nabla^2 H(x) v = -\sum_i \frac{v_i^2}{x_i}. \tag{54}$$

The key lower bound is

$$\sum_i \frac{v_i^2}{x_i} \geq 2\|v\|_2^2, \qquad v \in \mathcal{T}. \tag{55}$$

To prove it, write $P = \{i : v_i \geq 0\}$, $N = \{i : v_i < 0\}$, $a = \sum_{i \in P} v_i = -\sum_{i \in N} v_i$, and $s = \sum_{i \in P} x_i$. Cauchy–Schwarz gives

$$\sum_i \frac{v_i^2}{x_i} \geq \frac{a^2}{s} + \frac{a^2}{1-s} \geq 4a^2. \tag{56}$$

Since the positive and negative parts of v each have ℓ_1 mass a,

$$\|v\|_2^2 = \|v_+\|_2^2 + \|v_-\|_2^2 \leq a^2 + a^2 = 2a^2. \tag{57}$$

Combining the two inequalities yields (55). Equality holds for $n = 2$, $x = (1/2, 1/2)$, and $v = (1, -1)$.

Proof of the symmetric statement in Theorem 2. Write $\kappa = \lambda_{\max}(W_{\tan})$. For $x \in \Delta^\circ$ and $v \in \mathcal{T}$, (51) and (55) give

$$v^\top \nabla^2 \Phi_\beta(x) v = -\sum_i \frac{v_i^2}{x_i} + \beta v^\top W v \leq -2\|v\|_2^2 + \beta\kappa\|v\|_2^2 = -(2 - \beta\kappa)\|v\|_2^2. \tag{58}$$

If $\kappa \leq 0$, this is at most $-2\|v\|_2^2$; if $\kappa > 0$ and $\beta\kappa < 2$, it is strictly negative with modulus $2 - \beta\kappa$. With $\mu = 2 - \beta\kappa > 0$, integrate the Hessian inequality along an interior segment to obtain (20) for $x, y \in \Delta^\circ$. For arbitrary endpoints, set $x_\varepsilon = (1-\varepsilon)x + \varepsilon \mathbf{1}/n$ and $y_\varepsilon = (1-\varepsilon)y + \varepsilon \mathbf{1}/n$. These points are interior; continuity of the potential with $0 \log 0 = 0$ lets $\varepsilon \downarrow 0$ and proves the same strong-concavity inequality on the simplex on Δ. No entropy function on the entire affine hull is needed.

Compactness gives a global maximizer, the boundary argument above makes it interior, and strict concavity makes it unique. The KKT equivalence above identifies this maximizer with a logit fixed point. Conversely, any fixed point is an interior KKT point; for a concave differentiable function, such a point is a global maximizer because for every $y \in \Delta$,

$$\Phi_\beta(y) \leq \Phi_\beta(x) + \langle \nabla \Phi_\beta(x), y - x \rangle = \Phi_\beta(x). \tag{59}$$

Strict concavity therefore rules out any second fixed point. □

Several populations.

Proof of the block potential bound. On the relative interior of $\mathcal{X}$, the gradient of (22) in block a is

$$\nabla_{x^a} \mathcal{V}(x)_i = -\frac{1}{\beta_a}(1 + \log x_i^a) + (Wx + b)_i^a. \tag{60}$$

The KKT condition on each block is therefore

$$-\frac{1}{\beta_a}(1 + \log x_i^a) + (Wx + b)_i^a = \lambda_a, \tag{61}$$

which is equivalent, after exponentiating and normalizing within block a, to

$$x^a = \sigma_a\left(\beta_a (Wx + b)^a\right). \tag{62}$$

Thus fixed points of (12) are exactly interior KKT points of $\mathcal{V}$. As in the single-simplex case, the entropy term sends every maximizer into the relative interior of each block.

For $v \in \mathcal{T}$,

$$v^\top \nabla^2 \mathcal{V}(x) v = -\sum_{a=1}^{m} \frac{1}{\beta_a} \sum_i \frac{(v_i^a)^2}{x_i^a} + v^\top W v. \tag{63}$$

By applying (55) blockwise,

$$v^\top \nabla^2 \mathcal{V}(x) v \le -2 \sum_{a=1}^{m} \frac{1}{\beta_a} \|v^a\|_2^2 + v^\top W v. \tag{64}$$

Let $y = B_\beta^{-1/2} v$. Since B_β is scalar on each block, $y \in \mathcal{T}$ and $v = B_\beta^{1/2} y$. Hence

$$v^\top \nabla^2 \mathcal{V}(x) v \le -2\|y\|_2^2 + y^\top B_\beta^{1/2} \Pi W \Pi B_\beta^{1/2} y. \tag{65}$$

If the largest tangent eigenvalue of $B_\beta^{1/2} \Pi W \Pi B_\beta^{1/2}$ is strictly smaller than 2, the Hessian is strictly negative on every nonzero feasible direction. More precisely, with μ_{prod} as in (23),

$$v^\top \nabla^2 \mathcal{V}(x) v \le -\mu_{\mathrm{prod}} \|B_\beta^{-1/2} v\|_2^2.$$

Integration gives strong concavity on the simplex along interior feasible segments in this weighted norm. Approximating each endpoint by a mixture with the block-uniform interior point and passing to the limit using $0 \log 0 = 0$ proves the inequality on all of $\mathcal{X}$. The same compactness, interiority, and KKT argument used above gives a unique maximizer, and this maximizer is the unique block-logit fixed point. □

A.4 Convergence of learning dynamics

Sampled responses.

Proposition A.1. *Let $F : \Delta \to \Delta$ be a q-contraction in Euclidean norm, $q < 1$, with fixed point $x^\star$. Given $\mathcal{F}_k$, draw $A_{k+1}^{(j)}$, $j = 1, \dots, M$, independently with law $\mathrm{Cat}(F(x_k))$, and set*

$$\widehat{F}_{k+1} = \frac{1}{M} \sum_{j=1}^{M} e_{A_{k+1}^{(j)}}, \qquad x_{k+1} = (1 - \alpha_k) x_k + \alpha_k \widehat{F}_{k+1}.$$

For deterministic $0 < \alpha_k \le 1$, (25) holds. Robbins–Monro steps give almost-sure and mean-square convergence. If $\alpha_k = a/(k + k_0)$ with $k_0 \ge a > 0$ and $r = 2a(1 - q) > 1$, then

$$\mathbb{E}\|x_k - x^\star\|_2^2 \le C_0 (k + k_0)^{-r} + \frac{C_1}{M(k + k_0)}, \tag{66}$$

where constants depend on a, k_0, q and the initial error, but not on M or k. Constant steps give (26).

Proof. The update is a convex combination of points in Δ, so it stays feasible. Put $\xi_{k+1} = \widehat{F}_{k+1} - F(x_k)$ and $c = 1 - q > 0$. Conditional on the history,

$$\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\xi_{k+1}\|_2^2 \mid \mathcal{F}_k] = \frac{1 - \|F(x_k)\|_2^2}{M} \le \frac{1}{M}.$$

The conditional mean update has distance at most $(1 - c\alpha_k)\|x_k - x^\star\|_2$ from $x^\star$. Expanding squared distance makes the cross term vanish and proves (25).

For completeness, set $V_k = \|x_k - x^\star\|_2^2$. Since $0 \le c\alpha_k \le 1$, $(1 - c\alpha_k)^2 \le 1 - c\alpha_k$. With square-summable steps, define $Z_k = V_k + M^{-1} \sum_{j=k}^{\infty} \alpha_j^2$. Then Z_k is a nonnegative supermartingale satisfying $\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \le Z_k - c\alpha_k V_k$. The nonnegative supermartingale convergence theorem gives a finite almost-sure limit. Telescoping expectations gives $\mathbb{E} \sum_k c\alpha_k V_k \le \mathbb{E} Z_0 < \infty$, so this sum is finite almost surely. Because the tail correction vanishes, V_k converges; since $\sum_k \alpha_k = \infty$, its limit must be zero. The simplex has bounded diameter, so dominated convergence also gives mean-square convergence.

For harmonic steps, unroll the expectation recursion. Its homogeneous products satisfy

$$\prod_{j=\ell}^{k-1} \left(1 - \frac{ac}{j + k_0}\right)^2 \le C \left(\frac{\ell + k_0}{k + k_0}\right)^{2ac},$$

using $1 - u \le e^{-u}$ and the harmonic-sum lower bound; any zero initial factor only strengthens the bound. The accumulated noise is consequently at most

$$\frac{Ca^2}{M(k+k_0)^r} \sum_{j=0}^{k-1} (j+k_0)^{r-2} \le \frac{C'}{M(k+k_0)} \qquad (r > 1).$$

The initial term decays as $(k+k_0)^{-r}$, proving (66).

For constant α, the expectation recursion is geometric with coefficient $(1-c\alpha)^2 < 1$. Its limiting bound is

$$\frac{\alpha^2/M}{1-(1-c\alpha)^2} = \frac{\alpha}{Mc(2-c\alpha)},$$

which proves (26). □

For m blocks sampled with M draws per block, the squared Euclidean noise is the sum of the block squared noises and has conditional expectation at most m/M. All statements hold with this variance factor. More generally it is at most $\sum_a 1/M_a$ for block-specific batch sizes. Cross-block independence is not needed for this squared-norm bound, but conditional unbiasedness and the within-block variance assumptions are needed.

With $M = 1$ and $\alpha_k = 1$, every iterate after the first is a vertex. Its distance to an interior $x^\star$ is bounded away from zero, making convergence to $x^\star$ impossible. The constant-step error bound still applies. Extending the analysis to biased or dependent samples requires control of their conditional bias and variance.

Independent natural policy gradient.

Proof of Theorem 3. For a finite static game, each G_i is multilinear in the other players' mixed strategies and independent of π_i. It extends smoothly to a neighborhood of the product simplex. Positive policies stay positive under (27). Taking logarithms and applying Π_i removes each log-normalizing constant, giving (28) exactly. Conversely, for $s \in \mathcal{T}$, $\Pi \log \sigma_{\mathrm{blk}}(s) = s$, so this coordinate system parametrizes all full-support policies without redundant constants.

Let $\mathcal{H}(s) = (1-\alpha\tau)s + \alpha\Pi G(\sigma_{\mathrm{blk}}(s))$. The derivative on $\mathcal{T}$ is

$$D\mathcal{H}(s)|_{\mathcal{T}} = (1-\alpha\tau)I_{\mathcal{T}} + \alpha\Pi DG(\pi)\Pi\Sigma_{\mathrm{blk}}(\pi)|_{\mathcal{T}}. \tag{67}$$

The covariance matrix maps every perturbation into the tangent space. Hence, for $0 < \alpha\tau \le 1$, $\|D\mathcal{H}(s)|_{\mathcal{T}}\|_2 \le 1 - \alpha\tau + \alpha L_T/2 = \rho$. The bound holds uniformly on the complete vector space $\mathcal{T}$, so integration along score segments and Banach's theorem give a unique $s^\star$ and global linear convergence from any finite s_0.

At a fixed point, $\tau s^\star = \Pi G(\pi^\star)$; exponentiation gives $\pi_i^\star = \sigma_i(G_i(\pi^\star)/\tau)$. These are precisely the entropy-regularized best-response equations, so $\pi^\star$ is a QRE. Every QRE has full support when $\tau > 0$ and finite rewards, and thus corresponds to a fixed point in $\mathcal{T}$, proving uniqueness. The softmax $1/2$ bound turns the score rate into the stated policy rate. □

A bound in terms of the number of actions. For $i \ne j$, the (a_i, a_j) entry of $D_j G_i$ is

$$(D_j G_i)_{a_i,a_j} = \sum_{a_{-\{i,j\}}} r_i(a_i, a_j, a_{-\{i,j\}}) \prod_{\ell \notin \{i,j\}} \pi_\ell(a_\ell). \tag{68}$$

Under $r_i \in [0,1]$, every entry belongs to $[0,1]$. For a row of length n_j, centering subtracts its mean and leaves squared norm at most $n_j/4$, by the variance bound for numbers in $[0,1]$. Left projection cannot increase the Frobenius norm, so

$$\|\Pi_i D_j G_i \Pi_j\|_F^2 \le n_i n_j/4.$$

Diagonal derivative blocks vanish. Since the spectral norm is at most the Frobenius norm,

$$L_T^2 \le \frac{1}{4}\sum_{i\ne j} n_i n_j = \frac{1}{4}\left(S^2 - \sum_i n_i^2\right), \qquad S = \sum_i n_i. \tag{69}$$

Thus the sufficient lower bound on the entropy-regularization coefficient is $\frac{1}{4}\sqrt{S^2 - \sum_i n_i^2} < S/4 = (2S)/8$. The two results use different error measures and step-size conditions: Theorem III.3 of Sun et al. (2024) bounds QRE gap under $\tau > 2S$, whereas Theorem 3 bounds Euclidean distance between centered log policies under $0 < \alpha\tau \leq 1$. Both assume exact evaluation in finite static games.

For a polymatrix game, $G_i(\pi) = b_i + \sum_{j \neq i} W^{ij}\pi_j$ and $L_T = \|\Pi W \Pi|_{\mathcal{T}}\|_2$ can be computed directly. In a general game, the displayed reward bound or another verified uniform upper bound is sufficient. Projected Jacobian-vector and vector-Jacobian products can estimate singular values numerically, but finitely many trajectory measurements do not certify a global supremum.

B Additional results

This appendix extends the convergence argument to other norms and regularizers, compares it with a Dobrushin-type block contraction condition, and derives the applications summarized in Section 3.4. Each result is followed by its proof or derivation.

B.1 Other norms

The quotient space $\mathbb{R}^n/\operatorname{span}\{\mathbf{1}\}$ identifies payoff vectors that differ by an additive constant shared by all actions. In Euclidean geometry, Πz is its representative of smallest norm. Other norms require their corresponding quotient norm. Let $\mathcal{Q}_p = \mathbb{R}^n/\operatorname{span}\{\mathbf{1}\}$ and $\|[u]\|_{p,\mathrm{quot}} = \inf_{c\in\mathbb{R}} \|u + c\mathbf{1}\|_p$. The induced linear map is $\bar{W}h = [Wh]$ for $h \in \mathcal{T}$.

Proposition B.1. *For $1 \leq p \leq \infty$, softmax induces a $1/2$-Lipschitz map from $\mathcal{Q}_p$ to $(\Delta, \|\cdot\|_p)$. Therefore*

$$\mathrm{Lip}_p(F_\beta) \leq \tfrac{\beta}{2}\|\bar{W}\|_{(\mathcal{T},\|\cdot\|_p)\to\mathcal{Q}_p}.$$

If this bound is less than one, the map has a unique fixed point and globally convergent Picard iteration. For each p, no constant larger than 2 works for all systems under a condition based only on this operator norm.

Proof. Every absolute row and column sum of $\Sigma(p)$ is $2p_i(1-p_i) \leq 1/2$. Hence its induced 1- and ∞-norms are at most $1/2$; interpolation gives the same bound for every intermediate p, as in Nair (2026). Integrating the derivative gives $\|\sigma(u) - \sigma(v)\|_p \leq \|u - v\|_p/2$. For any constant c, invariance allows replacing $u - v$ by $u - v + c\mathbf{1}$ without changing the left side. Taking the infimum proves the quotient estimate and then the bound for the composed response.

For the binary example, a tangent vector is $(t, -t)$ and its image under W is the same vector. Symmetry and convexity imply that $c = 0$ minimizes $\|(t+c, -t+c)\|_p$ (also for $p = \infty$). Thus $\|\bar{W}\| = 1$ in every one of these geometries. The derivative at the uniform distribution has operator norm $\beta/2$, and there are multiple fixed points when $\beta > 2$. No larger constant can therefore guarantee either contraction or uniqueness for all systems using only this norm. □

For $p = 2$, the minimum-norm representative is Πu, recovering $W_{\tan}$. For other norms this identification generally fails: for example, for $u = (1, 0, 0)$ the ℓ_∞ quotient minimum is $1/2$, whereas $\|\Pi u\|_\infty = 2/3$. Retaining the quotient is essential to an intrinsic formulation.

B.2 Strongly convex regularizers

Fix a primal norm $\|\cdot\|$ on the ambient finite-dimensional space, with dual norm $\|\cdot\|_*$. Give payoff classes the quotient norm $\|[y]\|_{*,\mathrm{quot}} = \inf_c \|y + c\mathbf{1}\|_*$. Suppose $h : \Delta \to \mathbb{R} \cup \{+\infty\}$ is proper, lower-semicontinuous, and μ_h-strongly convex on its effective domain along feasible segments, with $\mu_h > 0$. This allows boundary-valued responses and does not assume an interior optimizer.

Proposition B.2. *The response $Q_h(y) = \arg\max_{x\in\Delta}\{\langle y, x\rangle - h(x)\}$ exists, is unique, is invariant to adding $c\mathbf{1}$, and satisfies*

$$\|Q_h(y) - Q_h(z)\| \leq \frac{1}{\mu_h}\|[y - z]\|_{*,\mathrm{quot}}. \tag{70}$$

Consequently, $x \mapsto Q_h(\beta(Wx + b))$ is contractive if $\beta\|\bar{W}\|_{(\mathcal{T},\|\cdot\|)\to(\mathcal{Q},\|\cdot\|_{*,\mathrm{quot}})} < \mu_h$.

Proof. Lower semicontinuity on compact Δ gives existence; strong convexity gives uniqueness. Let $x = Q_h(y)$ and $x' = Q_h(z)$. Applying strong convexity to the objective $h(\cdot) - \langle y, \cdot \rangle$ along the segment from its minimizer x to x' and taking a one-sided limit gives

$$h(x') \geq h(x) + \langle y, x' - x \rangle + \tfrac{\mu_h}{2}\|x - x'\|^2.$$

The analogous inequality with x, x' and y, z interchanged implies

$$\mu_h\|x - x'\|^2 \leq \langle y - z, x - x' \rangle \leq \|y - z + c\mathbf{1}\|_*\|x - x'\|$$

for every c, because $x - x' \in \mathcal{T}$. Taking the infimum and dividing, unless $x = x'$, proves (70). Invariance follows because a constant payoff shifts the maximized objective by a constant. Composing with the induced payoff map proves contraction on the complete simplex. □

For $h = -H$, Appendix A.3 proves strong convexity with Euclidean modulus 2, including boundary endpoints. At the uniform binary point the tangent Hessian has eigenvalue 2, so a larger dimension-free modulus is impossible. A Tsallis regularizer or another mirror map fits Proposition B.2 only after verifying a uniform positive modulus in the chosen geometry. Its boundary response, best modulus, and sharpness must be analyzed separately.

B.3 Comparison with a Dobrushin-type condition

A matrix of blockwise Lipschitz bounds takes the norm of each block before combining their effects. The Euclidean operator norm combines signed interactions first, allowing opposing effects to cancel. The following construction makes this distinction explicit.

For inverse temperatures β_a, the block influence matrix is

$$C_{ab} = \frac{\beta_a}{2}\|\Pi_a W^{ab} \Pi_b\|_{1\to 1}.$$

Here C_{ab} bounds the change in population a's response per unit change in population b, measured in ℓ_1. The Dobrushin-type block contraction criterion associated with this C is $\rho(C) < 1$, where $\rho(C)$ is the spectral radius. It uses the specified blockwise Lipschitz bounds, as in nonnegative-matrix convergence arguments (Mooij & Kappen, 2007). A strictly upper-triangular C has spectral radius zero even if some entries are large, so this condition can hold when the Euclidean norm bound fails. The example below gives the reverse separation.

Proposition B.3. *For every power of two m and every $\alpha \in (2/\sqrt{m}, 2)$, there exists a logit-response system with m two-action populations and unit inverse temperatures such that*

$$\frac{1}{2}\|\Pi W \Pi\|_{\mathcal{T}\to\mathcal{T}} = \frac{\alpha}{2} < 1, \tag{71}$$

so the condition in (14) *certifies global uniqueness and convergence, while the Dobrushin-type block ℓ_1 condition based on*

$$C_{ab} = \frac{1}{2}\|\Pi_a W^{ab} \Pi_b\|_{1\to 1} \tag{72}$$

fails because

$$\rho(C) = \frac{\alpha\sqrt{m}}{2} > 1. \tag{73}$$

Taking $\alpha = 1$ and $m \geq 8$ gives a system whose Euclidean contraction factor is $1/2$, while the spectral radius of the block ℓ_1 influence matrix already exceeds one. Figure 7 shows how this difference grows with the number of populations.

Proof of Proposition B.3. Let every block be binary, so $\Delta_a = \Delta_2$ and $\mathcal{T}_a = \operatorname{span}\{u\}$ with

$$u = \frac{1}{\sqrt{2}}(1, -1)^\top. \tag{74}$$

For every power of two m, take a Sylvester Hadamard matrix $H_m \in \{\pm 1\}^{m\times m}$, so that $m^{-1/2}H_m$ is orthogonal. Define the block interaction matrix by

$$W^{ab} = \frac{\alpha}{\sqrt{m}}(H_m)_{ab} u u^\top, \qquad a, b \in [m], \tag{75}$$

and take all inverse temperatures equal to one. Since $u \in \mathcal{T}_a$ and $uu^\top$ annihilates the all-ones direction, $\Pi_a W^{ab} \Pi_b = W^{ab}$.

Represent a tangent vector $h \in \mathcal{T}$ as $h^b = t_b u$. Then

$$(\Pi W \Pi h)^a = \sum_{b=1}^{m} \frac{\alpha}{\sqrt{m}} (H_m)_{ab} u u^\top (t_b u) = \left(\sum_{b=1}^{m} \frac{\alpha}{\sqrt{m}} (H_m)_{ab} t_b \right) u. \tag{76}$$

Thus the tangent operator is represented in the orthonormal block basis by

$$A = \frac{\alpha}{\sqrt{m}} H_m, \tag{77}$$

so $\|\Pi W \Pi\|_{\mathcal{T} \to \mathcal{T}} = \|A\|_2 = \alpha$. Hence

$$\frac{1}{2} \|\Pi W \Pi\|_{\mathcal{T} \to \mathcal{T}} = \frac{\alpha}{2} < 1. \tag{78}$$

The block contraction bound therefore certifies global uniqueness and convergence.

For the block ℓ_1 influence matrix, note that

$$u u^\top = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \qquad \|u u^\top\|_{1 \to 1} = 1. \tag{79}$$

Therefore

$$C_{ab} = \frac{1}{2} \|W^{ab}\|_{1 \to 1} = \frac{\alpha}{2\sqrt{m}} \tag{80}$$

for all a, b. Hence C is the all-ones matrix multiplied by $\alpha/(2\sqrt{m})$, and

$$\rho(C) = m \cdot \frac{\alpha}{2\sqrt{m}} = \frac{\alpha \sqrt{m}}{2} > 1 \tag{81}$$

by the assumption $\alpha > 2/\sqrt{m}$. The Dobrushin-type block ℓ_1 condition therefore fails. □

The role of the softmax constant. The improvement from 1 to 2 is most visible in the interval

$$1 \leq \beta \|W_{\tan}\|_2 < 2. \tag{82}$$

A unit Lipschitz estimate for softmax cannot certify this region. Using the sharp constant guarantees uniqueness and global convergence throughout it, for both Picard iteration and continuous adjustment. Thus the sufficient range for inverse temperature, or rationality, doubles relative to that coarse estimate on the same projected payoff matrix. In the binary example, this larger range reaches the bifurcation point $\beta = 2$.

B.4 Population and score dynamics

Population logit adjustment. For a unit-mass population with affine rewards, consider

$$\dot{x} = \sigma((Wx + b)/\eta) - x, \qquad \eta > 0.$$

Theorem 1, with $\beta = 1/\eta$, gives a unique globally exponentially attracting logit equilibrium when $\eta > \|W_{\tan}\|_2/2$, with exponent $1 - \|W_{\tan}\|_2/(2\eta)$. A Euclidean Lipschitz constant L_{pay} for $x \mapsto Wx + b$ on Δ satisfies

$$\|W_{\tan}\|_2 \leq \|W|_{\mathcal{T}}\|_{\mathcal{T} \to \mathbb{R}^n} \leq L_{\text{pay}}.$$

The second inequality follows because every tangent direction can be realized by a sufficiently small displacement from an interior point. This yields the stated comparison with the Lipschitz/passivity sufficient condition discussed by Cianfanelli & Como (2025). Structural results valid at every noise level rely on additional game assumptions. If population masses are $v_a > 0$, write $x^a = v_a y^a$ and absorb v_c into block W^{ac} before using the normalized product-simplex result; norms must be evaluated in these normalized coordinates.

Discounted score learning.

Corollary B.4. *For $\dot{z} = W\sigma(\beta z) + b - z$, suppose $q = \beta\|W_{\tan}\|_2/2 < 1$. Then the centered score $s = \Pi z$ and induced policy converge globally exponentially to unique limits. The full score z also converges to a unique limit.*

Proof. Payoff-shift invariance gives

$$\dot{s} = R(s) - s, \qquad R(s) = \Pi W\sigma(\beta s) + \Pi b.$$

For $h \in \mathcal{T}$, $DR(s)h = \beta\Pi W\Pi\Sigma(\sigma(\beta s))h$, whose norm is at most $q\|h\|_2$. Thus R contracts on the complete space $\mathcal{T}$, has a unique fixed point, and the same squared-distance argument as for logit adjustment gives exponent $1 - q$. The vector field has at most linear growth, so solutions exist globally. Softmax transfers the convergence to policies.

Write $z = s + c\mathbf{1}$, where $c = \mathbf{1}^\top z/n$. Then

$$\dot{c} = g(s) - c, \qquad g(s) = \mathbf{1}^\top(W\sigma(\beta s) + b)/n.$$

Variation of constants and $g(s(t)) \to g(s^\star)$ imply $c(t) \to g(s^\star)$, proving full-score convergence. For $q > 0$ an exponential forcing estimate yields the same exponent up to a constant; at $q = 0$ a te^{-t} term may appear for the common component. The full score converges in both cases. □

Adding $\mathbf{1}a^\top$ to W can inflate $\|W\Pi\|_2$ without changing $\Pi W\Pi$ or the induced policy dynamics. Centering the scores removes this irrelevant contribution from the convergence condition.

B.5 Mean-field iteration

Let $\mathcal{M}$ be a nonempty closed convex finite-dimensional domain of feasible mean-field flows, endowed with a Euclidean norm. Consider the exact finite-state, finite-horizon Boltzmann iteration of Cui & Koeppl (2021), restricted to a subclass for which $Q^\star(\mu)$ is C^1 on a neighborhood of $\mathcal{M}$. Center every state–time action block using Π_A. Define the direction space of the affine hull of $\mathcal{M}$ by

$$\mathcal{T}_{\mathcal{M}} := \operatorname{span}\{\mu - \mu' : \mu, \mu' \in \mathcal{M}\}.$$

Every difference $\mu - \mu'$ belongs to this fixed linear space, so the derivative restriction below controls changes along segments in $\mathcal{M}$. Assume

$$L_Q = \sup_{\mu\in\mathcal{M}} \|\Pi_A DQ^\star(\mu)|_{\mathcal{T}_{\mathcal{M}}}\|_2 < \infty.$$

For a fixed full-support reference policy ν, put

$$\Phi_\eta(\mu) = \sigma_{\mathrm{blk}}(Q^\star(\mu)/\eta + \log\nu).$$

Uniform ν gives ordinary Boltzmann policies. Let $\mathcal{P}$ be a $K_{\mathcal{P}}$-Lipschitz map from these policy blocks to mean-field flows in the specified compatible norms, and assume $\Gamma_\eta = \mathcal{P} \circ \Phi_\eta$ maps $\mathcal{M}$ to itself. The reference policy is fixed, so its logarithm contributes no derivative. The covariance bound and a segment integral give

$$\|\Phi_\eta(\mu) - \Phi_\eta(\mu')\|_2 \le \frac{L_Q}{2\eta}\|\mu - \mu'\|_2.$$

Composition gives

$$\mathrm{Lip}_2(\Gamma_\eta) \le \frac{K_{\mathcal{P}} L_Q}{2\eta}. \tag{83}$$

If $\eta > K_{\mathcal{P}} L_Q/2$, Banach's theorem gives a unique fixed point and convergence of exact iteration. If $Q^\star$ itself depends on η, L_Q must be evaluated at that temperature; the condition is then implicit.

The estimate applies to the softmax step of exact iteration in the stated Euclidean norms. Other steps and other metrics require their own bounds. It requires a fixed positive prior, the stated differentiability and Lipschitz assumptions, and exact maps. Optimal-value ties can invalidate differentiability; DQN approximation, particle error, and Markovian samples require separate treatment. As η decreases a Boltzmann policy becomes more concentrated on maximizing actions for fixed payoffs, but the resulting fixed point and the associated policy exploitability need not improve monotonically.

C EXPERIMENTAL DETAILS

We give the configurations and measurement procedures for Section 4, followed by numerical illustrations of sharpness and the comparison of sufficient conditions. All computations use double precision and synthetic data.

C.1 INVARIANCE TO PAYOFF SHIFTS

Let $Q = [(1,-1,0)^\top/\sqrt{2},\ (1,1,-2)^\top/\sqrt{6}]$ be an orthonormal basis for the three-action simplex tangent space, and let $R = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Set $W_0 = 1.6QRQ^\top$ and $W_c = W_0 + c\mathbf{1}Q_{:,1}^\top$. Because $Q^\top\mathbf{1} = 0$,

$$Q^\top W_c Q = 1.6R, \qquad W_c^\top W_c\big|_{\mathcal{T}} \simeq 1.6^2 I_2 + 3c^2 e_1 e_1^\top.$$

Thus $\|W_c\|_2 = \sqrt{1.6^2 + 3c^2}$ while the projected matrix norm remains 1.6 and the softmax response is unchanged. We sweep c through 121 equally spaced values in $[0, 12]$. Choosing $c = \sqrt{(16.75^2 - 1.6^2)/3}$ gives full-matrix norm 16.75.

We use the three vertices and 61 Dirichlet$(1,1,1)$ initializations from seed 20260925, with $\beta = 1$ and $b = 0$. The fixed point is $\mathbf{1}/3$. The maximum componentwise difference between shifted and unshifted responses at these initializations is 1.67×10^{-16}. The largest observed error ratio is 0.5803, below the contraction factor 0.8. For each initialization j, let $e_{j,k} = \|x_j^k - x^\star\|_2$. The reported statistic is

$$\max_{\substack{j,\ 0\le k\le 60 \\ e_{j,k} > 10^{-10}}} \frac{e_{j,k+1}}{e_{j,k}}.$$

The cutoff avoids ratios dominated by floating-point noise. We save errors for $k = 0, \ldots, 60$ and evaluate the next response at each step to compute the ratio. Figure 1 plots the maximum distance and the theoretical error bound.

C.2 NONLINEAR AND POPULATION DYNAMICS

For the nonlinear example, use

$$\begin{aligned} G(x) &= 1.1QRQ^\top x + b + C\tanh(Ax + d), \\ C &= 0.4Q, \quad A = 0.8Q^\top, \quad b = (0.15, -0.05, -0.10)^\top, \quad d = (0.2, -0.1)^\top. \end{aligned} \tag{84}$$

At $\beta = 1$, (50) gives $L_T \le 1.1 + 0.4 \cdot 0.8 = 1.42$ and $q \le 0.71$. The unit Lipschitz bound using this same estimate cannot certify contraction. Using the same 64 initializations, Picard iteration has largest observed error ratio 0.4179 and stays below the theoretical error bound. Its reference fixed point, found by iterating to a change below 10^{-15}, is approximately $(0.40085654, 0.31191615, 0.28722731)$; the contraction residual bound controls its numerical error.

We also instantiate the generalized rock–scissor–paper matrix in Example 2 of Cianfanelli & Como (2025), at their parameter $\gamma = 2$:

$$W_{\mathrm{RSP}} = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 2 & -1 \\ -1 & 1 & 2 \end{pmatrix}, \qquad b = 0, \qquad \eta = 1.5. \tag{85}$$

We choose the temperature $\eta = 1.5$ for this check. On the tangent plane, the symmetric part is $2I$ and the skew part has squared magnitude 3, so $\|Q^\top W_{\mathrm{RSP}} Q\|_2 = \sqrt{7}$. Thus the bound from a unit Lipschitz estimate is $\sqrt{7}/1.5 > 1$, while our Lipschitz bound is $\sqrt{7}/3 \approx 0.8819$. Picard iteration from the same starts converges to $\mathbf{1}/3$, with largest observed error ratio 0.6403 under the same cutoff. The observed trajectories satisfy the analytic contraction bounds.

C.3 STOCHASTIC RESPONSE UPDATES

Use the binary example $W = \left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$, $b = 0$, $\beta = 1.6$, and $q = 0.8$. Writing $u = x_1$,

$$F_1(x) = \mathrm{logistic}(1.6(2u - 1)), \qquad x^\star = (1/2, 1/2).$$

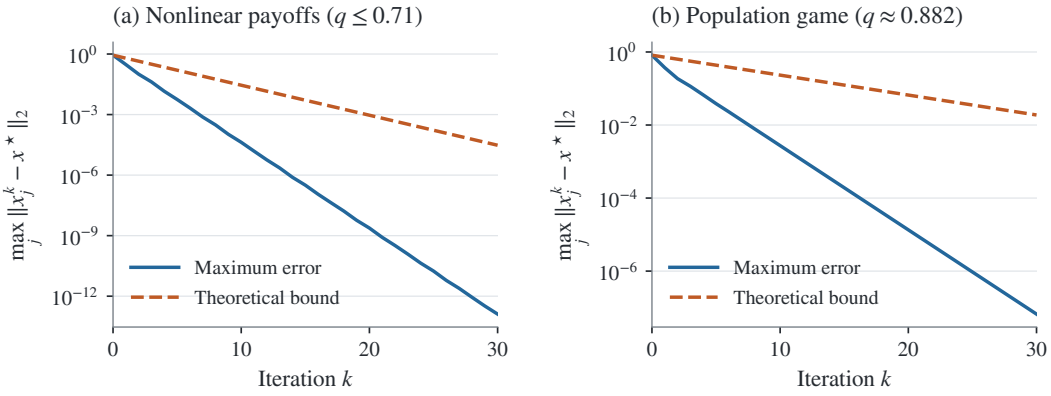


Figure 3: The sharp response bound certifies both examples where the unit Lipschitz estimate exceeds one. (a) Nonlinear payoffs: $q \le 0.71$, with largest observed error ratio 0.4179. (b) Generalized rock–scissor–paper (RSP) dynamics: $q \approx 0.8819$, with largest observed error ratio 0.6403. Solid curves show maximum error over 64 initializations; dashed curves give the analytic envelopes. Ratios exclude errors below 10^{-10}; plots show the first 30 steps.

At each update, sample the binomial count $B \sim \mathrm{Binomial}(M, F_1(x))$ and use $(B/M, 1 - B/M)$ as the empirical response. This is exactly the mean of M categorical samples. We run $M \in \{1, 4, 16, 64\}$, 512 trajectories for each setting, and 30,000 updates from $x_0 = (0.9, 0.1)$. The three seeds are 20260925, 20260926, and 20260927; NumPy PCG64 generators use SeedSequence([seed, mode index]) with indices zero and one for the two schedules.

The harmonic schedule is $\alpha_k = 5/(k + 50)$, for which $2a(1 - q) = 2 > 1$, and the reported MSE is the average across trajectories at the final step. The constant schedule is $\alpha = 0.05$, with MSE averaged across trajectories and the last 3,000 steps. For each seed and schedule, we fit an ordinary least-squares slope of $\log(\mathrm{MSE})$ against $\log M$ using all four batch sizes. Table 2 reports means and sample standard deviations across the three independent seeds.

Table 2: Sampling error and its theoretical bound. Entries are MSE mean $\pm$ sample standard deviation over three seeds, each with 512 trajectories. Diminishing-step MSE is measured at iteration 30,000; constant-step MSE is averaged over the final 3,000 iterations. The bound applies to constant steps. Column multipliers indicate the scale of each entry.

	Empirical MSE		Theoretical bound
Batch size M	Diminishing steps ($\times 10^{-4}$)	Constant steps ($\times 10^{-2}$)	Constant steps ($\times 10^{-2}$)
1	4.332 ± 0.244	4.983 ± 0.028	12.563
4	1.060 ± 0.092	1.443 ± 0.015	3.141
16	0.254 ± 0.004	0.386 ± 0.007	0.785
64	0.074 ± 0.005	0.097 ± 0.002	0.196

Table 2 shows that increasing the batch size reduces MSE under both schedules. The fitted slopes are -0.9844 ± 0.0214 for harmonic steps and -0.9466 ± 0.0018 for constant steps. The harmonic schedule closely follows inverse-batch-size scaling, while the constant-step slope is above -1. Every empirical constant-step tail MSE is below (26), with maximum empirical-to-bound ratio 0.5046.

The role of averaging is also apparent when $M = 1$ and $\alpha = 1$: every post-update state is a vertex and its squared distance to $(1/2, 1/2)$ is exactly $1/2$. This identity holds for every sample path.

Numerical implementation. The reported computations use Python 3.14.6, NumPy 2.5.1, SciPy 1.18.0, and Matplotlib 3.11.1. Errors are computed in Euclidean norm; the matrix norms in the convergence conditions are spectral norms.

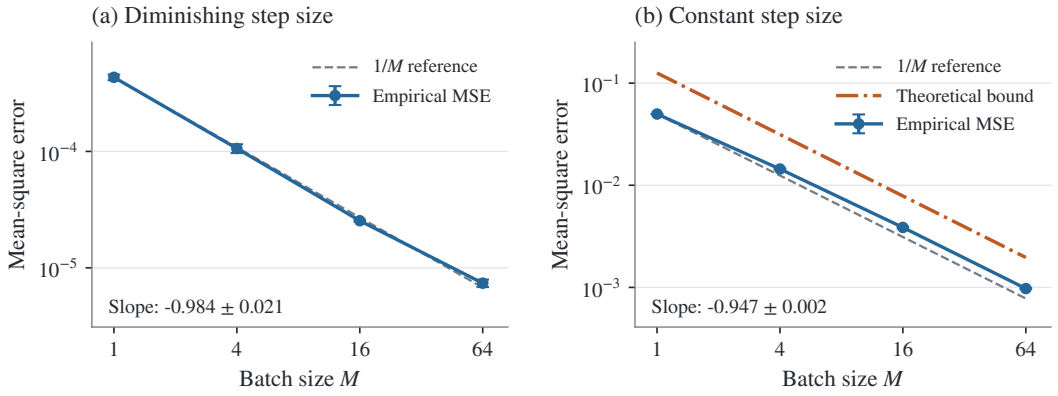


Figure 4: Measured sampling error decreases approximately as $1/M$ under both step schedules. Error bars show sample standard deviations across three seeds, each with 512 trajectories. The dashed line is a $1/M$ reference anchored at $M = 1$; the dash-dotted line is the constant-step expectation bound. All measured constant-step tail errors lie below this bound.

C.4 Additional numerical illustrations

These examples illustrate three distinct features of the condition: the binary model attains its universal boundary, projection enlarges the range where the sufficient condition holds in affine systems, and signed interactions allow cancellation across populations. The constructions use fixed seeds and no external data.

The two-action example. The binary model shows that the universal constant 2 coincides with an actual loss of uniqueness. Figure 2 plots the scalar equation in the two-action example in Section 3.1 together with continuous adjustment below and above the bifurcation.

For the nonzero equilibrium branches, we solve $m = \tanh(\beta m/2)$ by 120 bisection steps on $[10^{-10}, 1 - 10^{-10}]$, at 300 equally spaced values of β in $[2.0005, 3.35]$. The negative branch follows by symmetry. For the trajectories, we use $\beta \in \{1.5, 2.3\}$ and initial probability differences $m_0 = x_1(0) - x_2(0) \in \{-0.9, -0.45, 0.45, 0.9\}$. We integrate $\dot{m} = \tanh(\beta m/2) - m$ by the classical fourth-order Runge–Kutta method on 620 equally spaced time points in $[0, 14]$.

The additional interval of guaranteed convergence $1 \le \beta < 2$ in Figure 2 retains a unique global attractor. Above 2, the sign of the initial difference determines which of the two stable asymmetric equilibria is reached. The change in behavior explains why the universal constant in the sufficient norm condition cannot be increased.

Random affine interactions. We generate four ensembles of 32×32 matrices, with 200 draws per ensemble and a NumPy PCG64 generator initialized with seed 20260506. Let the entries of Z be independent $N(0, 1/32)$ variables. The unshifted nonsymmetric and symmetric matrices are Z and $(Z + Z^\top)/2$, respectively. For a shifted matrix, project the corresponding matrix on both sides by Π to obtain $W_{\tan}$. Draw independent standard Gaussian vectors and normalize each to unit Euclidean norm. Multiply them by 1.5 to obtain a and b in the nonsymmetric construction below; in the symmetric construction use the same a on both sides. The ensembles are generated in the order symmetric unshifted, nonsymmetric unshifted, symmetric shifted, and nonsymmetric shifted.

For a single-simplex affine logit system, the full-matrix and projected conditions guarantee contraction for $\beta < \beta_{\mathrm{ref}}$ and $\beta < \beta_{\mathrm{sharp}}$, respectively, where

$$\beta_{\mathrm{ref}} = \frac{1}{\|W\|_2}, \qquad \beta_{\mathrm{sharp}} = \frac{2}{\|\Pi W \Pi\|_{\mathcal{T}\to\mathcal{T}}}. \tag{86}$$

We report the ratio of these upper bounds

$$\frac{\beta_{\mathrm{sharp}}}{\beta_{\mathrm{ref}}} = \frac{2\|W\|_2}{\|\Pi W \Pi\|_{\mathcal{T}\to\mathcal{T}}}, \tag{87}$$

or equivalently its reciprocal $\beta_{\mathrm{ref}}/\beta_{\mathrm{sharp}}$. For ordinary Gaussian matrices, the projected matrix norm is usually close to the full-matrix norm, so the visible gain is primarily the factor two. To isolate the geometric correction, we also generate payoff-shift components of the form

$$W = W_{\tan} + \mathbf{1}a^\top + b\mathbf{1}^\top, \tag{88}$$

with the symmetric variant $W = W_{\tan} + \mathbf{1}a^\top + a\mathbf{1}^\top$. Since

$$\Pi(\mathbf{1}a^\top + b\mathbf{1}^\top)\Pi = 0, \tag{89}$$

these components can inflate $\|W\|_2$ without changing the projected payoff matrix.

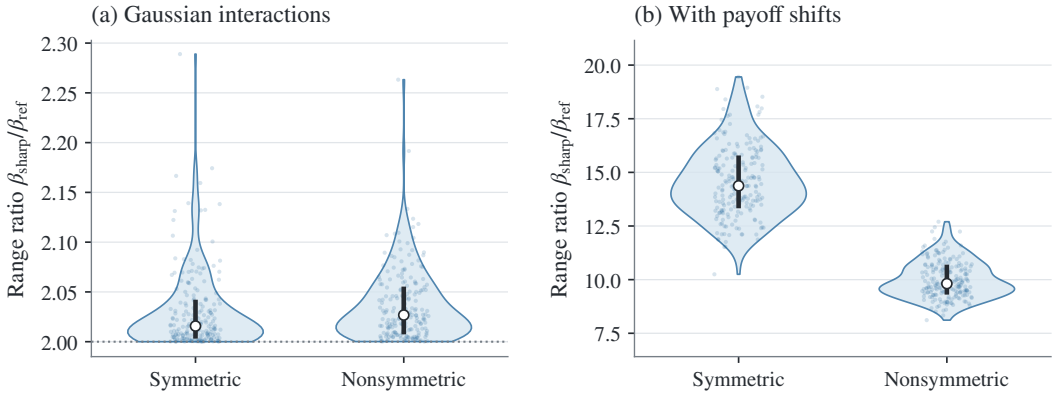


Figure 5: **Comparison of sufficient bounds for random affine logit systems.** Each distribution contains 200 draws with $n = 32$; points show individual draws, white markers show medians, and dark bars show interquartile ranges. The ordinate is $\beta_{\mathrm{sharp}}/\beta_{\mathrm{ref}} = 2\|W\|_2/\|\Pi W\Pi\|_{\mathcal{T}\to\mathcal{T}}$. (a) Gaussian interactions: the dotted line marks a factor of two. (b) Adding payoff shifts yields a larger improvement. The panels use different ordinate ranges to resolve the distributions.

Convergence of repeated responses. The next illustration checks nontrivial instances in which the coarse bound fails but the projected bound certifies contraction. We generate 12 non-symmetric shifted Gaussian systems with $n = 20$, random biases, and choose

$$\beta = 0.72\,\beta_{\mathrm{sharp}}. \tag{90}$$

Because $\beta_{\mathrm{sharp}}/\beta_{\mathrm{ref}} \geq 2$, this choice satisfies

$$\beta_{\mathrm{ref}} < \beta < \beta_{\mathrm{sharp}} \tag{91}$$

for every displayed instance. Thus the condition obtained from a unit softmax Lipschitz bound fails, while Theorem 1 certifies contraction. We use the nonsymmetric shifted construction above with $n = 20$, multiplying the normalized shift vectors by 1.4. Bias coordinates are independent $N(0, 0.2^2)$ variables. A PCG64 generator with seed 20260507 draws each matrix, its bias, and ten independent Dirichlet$(\mathbf{1}_{20})$ initializations in sequence. We run 90 Picard steps for each instance and plot the maximum pairwise distance among the iterates; Figure 6 shows the first twelve steps.

Comparison with a Dobrushin-type condition. The block contraction bound retains the signs of interactions when bounding their combined effect. The Hadamard block construction in Proposition B.3 gives, for $\alpha = 1$,

$$q_{\ell_2} = \frac{1}{2}\|\Pi W\Pi\|_{\mathcal{T}\to\mathcal{T}} = \frac{1}{2}, \qquad \rho(C) = \frac{\sqrt{m}}{2}. \tag{92}$$

Thus the Euclidean contraction condition on the product simplex remains valid as the number of binary blocks grows, while the Dobrushin-type block ℓ_1 condition fails once $m \geq 8$.

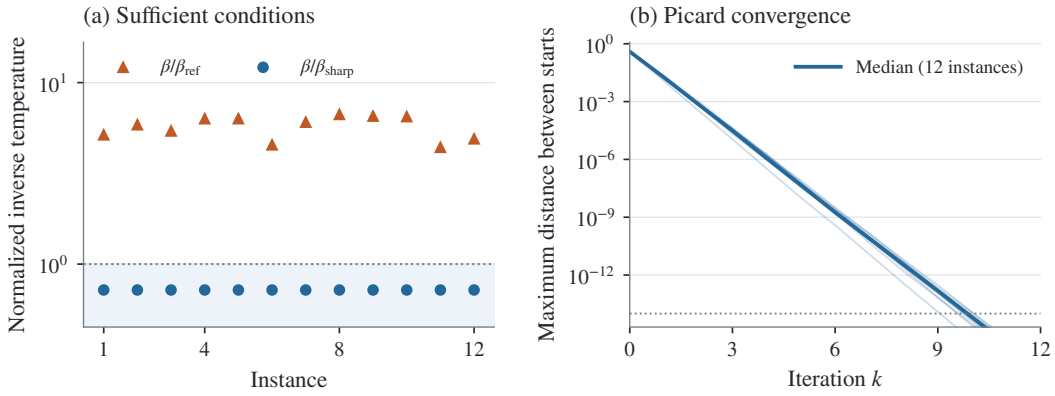


Figure 6: **Picard convergence under the projected norm condition.** Panel (a) compares the chosen β with the reference and sharp bounds for each instance. Panel (b) shows the maximum pairwise distance among ten initializations for each of twelve instances over the first twelve iterations. Thin curves represent individual instances, the thick curve is their pointwise median, and the dotted line marks 10^{-14}.

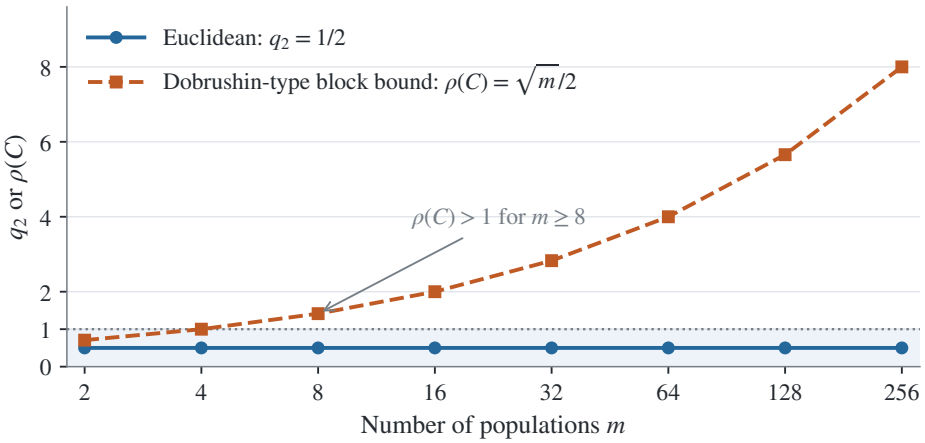


Figure 7: **Comparison of Euclidean and Dobrushin-type block conditions.** For the Hadamard block construction with $\alpha = 1$, the Euclidean contraction factor remains $q_{\ell_2} = 1/2$ for all powers-of-two block counts, while the spectral radius of the block ℓ_1 influence matrix grows as $\rho(C) = \sqrt{m}/2$ and exceeds one once $m \geq 8$.

D DISCUSSION

Interpretation of sharpness. The sharpness statements concern four distinct quantities. The softmax covariance satisfies $\sup_p \|\Sigma(p)\|_2 = 1/2$, with equality on a two-point distribution. The constant 2 is optimal among sufficient conditions of the form $\beta\|W_{\tan}\|_2 < C$ that apply to every payoff matrix: for any $C > 2$, the binary example has multiple fixed points at some parameter strictly between 2 and C. In the symmetric class, 2 is also the optimal entropy curvature constant along feasible directions. The binary family realizes this value as an actual bifurcation: there is one fixed point below 2 and three above it. Particular payoff matrices may admit uniqueness beyond the range guaranteed by the norm condition.

Limitations. The results concern finite-dimensional logit response maps. The affine condition is optimal among those based only on the projected matrix norm, but need not be necessary for each matrix. Bounds that account for how covariance varies across states, or restrict to an invariant region, may give stronger conclusions. The nonlinear extension requires a uniform bound on the projected Jacobian. Nonsmooth payoffs and infinite-dimensional systems require further analysis. General regularizers need their own convexity modulus, treatment of the boundary, and optimality argument. The Euclidean condition and the Dobrushin-type block condition based on the specified influence matrix are complementary; neither uniformly dominates the other.

The sampling analysis uses the conditional unbiasedness and variance assumptions in Appendix A.4. Biased or dependent observations require control of the resulting conditional error. With one sample and no averaging over time, the iterates remain at simplex vertices and cannot converge to an interior fixed point. Independent NPG assumes exact evaluation in a finite static game, while the mean-field application requires differentiable exact maps in compatible Euclidean norms. Extending these guarantees to function approximation, delayed updates, or particle estimates requires bounds on the corresponding approximation errors.

Broader impacts. This is foundational theoretical work with synthetic numerical examples; it uses no personal data and deploys no learned model. Its positive impact is to provide sharper sufficient conditions for stability and uniqueness in logit response models used in games, learning dynamics, and entropy-regularized optimization. Indirectly, such conditions may be used in deployed decision systems involving strategic populations or automated recommendations; in those settings, stability alone does not address fairness, privacy, welfare, or distributional harms. Any applied use should therefore be combined with domain-specific evaluation of these concerns.